Centro Federal de Educação Tecnológica de Minas Gerais
Departamento de Computação
Engenharia de Computação

Mariane Raquel Silva Gonçalves

Comparação de algoritmos paralelos de ordenação em MapReduce

Orientadora: Prof[a]. Dr[a]. Cristina Duarte Murta

Belo Horizonte
2012

1 Introdução

Fazer aqui uma introdução geral da área do conhecimento à qual o tema escolhido está ligado.

1.1 Definição do Problema

Na última década, a quantidade de dados *(de trabalho, utilizada pelos sistemas, disponíveis) elaborar mais* aumentou várias ordens de grandeza, fazendo do processamento dos dados um desafio para a computação sequencial. Como resultado, torna-se crucial substituir a computação tradicional por computação distribuída eficiente [Lin e Dyer 2010]. A mudança no modelo de programação sequencial para paralelo é um fato inevitável e ocorre gradualmente, desde que a indústria declarou que seu futuro está em computação paralela [Asanovic et al. 2009].

O MapReduce é um modelo de programação paralela desenvolvido pela Google para processamento de grandes volumes de dados distribuídos em *clusters* [Dean e Ghemawat 2008]. Esse modelo propõe simplificar a computação paralela, escondendo detalhes da paralelização do desenvolvedor e utilizando duas funções principais - map e reduce. Uma das implementações mais conhecidas e utilizadas do modelo é o Hadoop [White 2009], ferramenta de código aberto, desenvolvida por Doug Cutting em 2005 e apoiada pela Yahoo!.

A ordenação é um dos problemas fundamentais da ciência da computação e um dos problemas algorítmicos mais estudados. Muitas aplicações dependem de ordenações eficientes como base para seu próprio desempenho. A ordenação é um problema que abrange desde sistemas de banco de dados à computação gráfica, e muitos outros algoritmos podem ser descritos em termos de ordenação [Satish et al. 2009, Amato et al. 1998].

Uso crescente de computação paralela em sistemas computacionais gera a necessidade de algoritmos de ordenação inovadores, desenvolvidos para dar suporte a essas aplicações. Isso significa desenvolver rotinas eficientes de ordenação em

arquiteturas paralelas e distribuídas.

O trabalho proposto por Pinhão (2011) apresentou uma avaliação da escalabilidade de algoritmos de ordenação paralela no modelo MapReduce. Para tal, foi desenvolvido no ambiente Hadoop o algoritmo de Ordenação por Amostragem, e seu desempenho foi avaliado em relação à quantidade de dados de entrada e ao número de máquinas utilizadas.

Considerando esse contexto, o presente trabalho segue este tema e busca continuar a análise, com a implementação do algoritmo Quicksort no mesmo ambiente, bem como a análise de escalabilidade e comparação do desempenho dos algoritmos.

1.2 Motivação

O volume de dados que é produzido e tratado em indústrias, empresas e até mesmo em âmbito pessoal aumenta a cada ano. O desenvolvimento de soluções capazes de lidar com tais volumes de dados é uma das preocupações atuais, tendo em vista a quantidade de dados processados diariamente, e o rápido crescimento desse volume de dados. Não é fácil medir o volume total de dados armazenados digitamente, mas uma estimativa da IDC [Gantz 2008] calculou o tamanho do universo digital em 0,18 zettabytes em 2006, e previa um crescimento dez vezes até 2011 (chegando a 1,8 zettabytes). *The New York Stock Exchange* gera cerca de um terabyte de novos dados comerciais por dia. O Facebook armazena aproximadamente 10 bilhões de fotos, que ocupam mais de um petabyte. *The Internet Archive* armazena aproximadamente 2 petabytes de dados, com aumento de 20 terabytes por mês [White 2009]. Estima-se que dados não estruturados são a maior porção e a de mais rápido crescimento dentro das empresas, o que torna o processamento de tal volume de dados muitas vezes inviável.

Mesmo para os computadores atuais, é um desafio conseguir lidar com quantidades de dados tão grandes. É preciso buscar soluções escaláveis, que apresentem bom desempenho em tais condições. As tendências atuais em *design* de microprocessadores estão mudando fundamentalmente a maneira de se processar dados em sistemas computacionais.

Nos últimos 40 anos, o aumento no poder computacional deu-se, largamente, ao aumento na capacidade do hardware. O fator principal dessa melhoria foi a capacidade de dobrar, a cada dois anos, o número de dispositivos microeletrônicos em uma mesma área de silício, a um custo quase constante. Esse crescimento exponencial no número de transistores presentes nos processadores é conhecida como Lei de Moore [Manferdelli et al. 2008]. No entanto, o aumento no número de transistores, e consequente aumento na velocidade do processador foi limitado por questões físicas, como a dissipação do calor.

Há alguns anos a indústria percebeu tal fato, e declarou que seu futuro está em computação paralela, com o aumento crescente do número de núcleos dos processadores [Asanovic et al. 2009]. Atualmente, arquitetos sabem que o aumento no desempenho só pode ser alcançado com o uso de computação paralela, e têm recorrido cada vez mais a arquiteturas paralelas para continuar a fazer progressos [Manferdelli et al. 2008].

Com os novos modelos de processadores multicore, o modelo de programação single core está sendo substituído rapidamente pelo novo modelo, e com isso surge a necessidade de escrever software para sistemas com multiprocessadores e memória compartilhada [Ernst et al. 2009]. Arquiteturas multi-core podem oferecer um aumento significativo de desempenho sobre as arquiteturas de núcleo único, de acordo com as tarefas paralelizadas. No entanto, muitas vezes isto exige novos paradigmas de programação para utilizar eficientemente a arquitetura envolvida [Prinslow 2011].

A técnicas tradicionais de programação paralelas - como passagem de mensagens e memória compartilhada, em geral são complexas e de difícil entendimento para grande parte dos desenvolvedores. Em tais modelos, é preciso gerenciar localidades temporais e espaciais e lidar explicitamente com concorrência, criando e sincronizando *threads* através de mensagens e *semáforos*. Dessa forma, não é uma uma tarefa simples escrever códigos paralelos corretos e escaláveis para algoritmos não triviais [Ranger et al. 2007].

O MapReduce surgiu como uma alternativa aos modelos tradicionais, com o objetivo de simplificar a computação paralela. O maior benefício desse modelo é a simplicidade. O foco do programador é a descrição funcional do algoritmo, e não as formas de paralelização. Nos últimos anos o modelo têm se estabelecido

como uma das plataformas de computação paralela mais amplamente utilizadas no processamento de terabyte e petabyte de dados [Ranger et al. 2007]. MapReduce e sua implementação *open source* Hadoop oferecem uma alternativa economicamente atraente através de uma plataforma eficiente de computação distribuída, capaz de lidar com grandes volumes de dados e mineração de petabytes de informações não estruturadas [Cherkasova 2011].

// texto conector

A ordenação é um dos problemas fundamentais da ciência da computação e algoritmos paralelos para ordenação têm sido estudados desde o início da computação paralela. Os algoritmos ótimos existentes em arquitetura sequencial, como Quick Sort e Heap Sort necessitam de um tempo mínimo *(n log n)* para ordenar uma sequência de n elementos [Aho et al. 1974].

Na ordenação paralela, fatores como movimentação de dados, balanço de carga, latência de comunicação e distribuição inicial das chaves são considerados ingredientes chave para o bom desempenho, e variam de acordo com o algoritmo escolhido como solução [Kale e Solomonik 2010].

Dado o grande número de algoritmos de ordenação paralela e grande variedade de arquiteturas paralelas, é uma tarefa difícil escolher o melhor algoritmo para uma determinada máquina e instância do problema. Além disso, não existe um modelo teórico conhecido que pode ser aplicado para prever com precisão o desempenho de um algoritmo em arquiteturas diferentes [Amato et al. 1998].

Assim, estudos experimentais assumem uma crescente importância para a avaliação e seleção de algoritmos apropriados para multiprocessadores. É preciso que mais estudos sejam realizados para que determinado algoritmo pode ser recomendado em certa arquitetura com alto grau de confiança.

1.3 Objetivos

Os objetivos deste trabalho são:

- Estudar a programação paralela aplicada à algoritmos de ordenação;
- Implementar um ou mais algoritmos de ordenação paralela no modelo MapRe-

duce, com o software Hadoop;

- Comparar duas ou mais implementações de algoritmos paralelos de ordenação.

O trabalho desenvolvido por Pinhão (2011) apresentou um estudo sobre a computação paralela e algoritmos de ordenação no modelo MapReduce, através da implementação do algoritmo de Ordenação por Amostragem feita em ambiente Hadoop.

Este projeto busca continuar o estudo sobre ordenação paralela feito no trabalho citado, com a análise de desempenho dos algoritmos de ordenação ordenação por amostragem e quick sort. A análise busca compará-los com relação à quantidade de dados a serem ordenados, variabilidade dos dados de entrada e número máquinas utilizadas.

1.4 Organização do Texto

Esse projeto está organizado em cinco capítulos. O próximo capítulo apresenta o referencial teórico para o desenvolvimento do trabalho. O Capítulo 3 descreve a metologia de pesquisa, indicando os passos a serem seguidos durante o desenvolvimento. Os resultados preliminares obtidos até a entrega do projeto são apresentados no Capítulo 4. As conclusões obtidas até o momento e os próximos passos para a conclusão do projeto estão no Capítulo 5.

2 Referencial Teórico

2.1 Processamento Distribuído ? / Computação Paralela ?

Com o avanço tecnólogico da última década, o volume crescente de dados sendo gerado, coletado e armazenado tornou o processamento dos dados inviável para um único computador. A quantidade de dados atualmente processados cria a necessidade de computação de alto desempenho, cujo foco sejam os dados. Como resultado, torna-se crucial substituir a computação tradicional por computação distribuída eficiente. É um caminho natural para o processamento de dados em larga escala o uso de *clusters* [Lin e Dyer 2010].

Clusters são conjuntos de máquinas, ligadas em rede, que comunicam-se através do sistema, trabalhando como se fossem uma única máquina de grande porte. Dentre algumas características observadas em um *cluster*, é possível destacar: o baixo custo se comparado a supercomputadores; a proximidade geografica dos nós. altas taxas de transferência nas conexões entre as máquinas e o uso de máquinas homogêneas [**?**].

Apesar dos computadores em um *cluster* não precisarem processar necessariamente a mesma aplicação, a grande vantagem de tal organização é a habilidade de cada nó processar individualmente uma fração da aplicação, resultando em desempenho que pode ser comparado ao de um supercomputador. Em geral os computadores de *clusters* são de baixo custo, o que permite que um grande número de máquinas seja interligadas, garantindo grande desempenho e melhor custo-benefício que supercomputadores, o que apresenta grande vantagem. Outro ponto importante é que novas máquinas podem ser facilmente incorporadas ao *cluster*, tornando-o uma solução mais flexível, principalmente por ser formado por maquinas de capacidade de processamento similar.

Esta linha de pesquisa envolve vários outros conceitos relacionados à infraestrutura, como comunicação entre os nós, balanceamento de carga e outros discutidos nas próximas seções.

2.1.1 Princípios de processamento em cluster? distribuído? de grande volumes de dados?

O processamento de grandes volumes de dados em *clusters* deve suportar alguns princípios para garantir a escalabilidade e o bom desempenho [Bryant 2011]:

Tratamento de dados intrínsecos A coleta e manutenção dos dados deve ser funções do sistema e não tarefa dos usuários. O sistema deve recuperar informações atualizadas através de rede e realizar cálculos derivados como tarefas em segundo plano. Os usuários devem ser capazes de usar consultas ricos com base no conteúdo e identidade para acessar os dados. Mecanismos de confiabilidade, como replicação e de correção de erros devem ser incorporados como parte do sistema, de modo a garantir integridade e disponibilidade dos dados.

Modelo de programação paralelo de alto nível O desenvolvedor da aplicação deve fazer uso de primitivas de programação de alto nível, capazes de expressar formas naturais de paralelismo, que não incluam cofigurações específicas de uma máquina. O trabalho de Mapear essas computações para a máquina de forma eficiente deve ficar a cargo do sistema, compilador e runtime.

Acesso interativo Os usuários devem ser capazes de executar programas de forma interativa, com variação dos requisitos de computação e armazenamento O sistema deve responder a consultas e cálculos simples rapidamente, e responder aos complexos sem degradar o desempenho geral. Para suportar a computação interativa, deve haver oferta de recursos. O custo consequente do aumento dos recursos ofertados pode ser justificado com base no aumento da produtividade dos usuários do sistema.

Mecanismos escaláveis para garantir alta confiabilidade e disponibilidade Um sistema para computação de grandes volumes de dados deve implementar mecanismos de confiabilidade, no qual os dados originais e intermediários são armazenados de forma redundante. Isso permite que no caso de falhas de componente ou dados seja possível refazer a computação. Além disso, a máquina deve identificar e desativar automaticamente componentes que falharam, de modo a não prejudicar o

desempenho do sistema e se manter sempre disponível.

Grandes empresas de serviços de Internet - como Google, Yahoo, Facebook e Amazon - buscam soluções para processamento de dados em grandes conjuntos de máquinas que atendam as características descritas. Com um software que provê tais características é possível alcançar alto grau de escalabilidade e custo-desempenho.

Dentre as principais propostas está o modelo MapReduce e sua implementação Hadoop, que são soluções escaláveis, capazes de processar grandes volumes de dados, com alto nível de abstração para distribuir a aplicação e mecanismos de tolerância a falhas.

A próxima seção apresenta com mais detalhes o modelo e suas características.

2.2 MapReduce

O MapReduce é um modelo de programação paralela criado pela Google para processamento de grandes volumes de dados em *clusters*. Esse modelo propõe simplificar a computação paralela e ser de fácil uso, abstraindo conceitos complexos da paralelização - como tolerância a falhas, distribuição de dados e balanço de carga - e utilizando duas funções principais: Map e reduce. A complexidade do algoritmo paralelo não é vista pelo desenvolvedor, que pode se ocupar em desenvolver a solução proposta [Dean e Ghemawat 2008].

Esse modelo de programação é inspirado em linguagens funcionais, tendo como base as primitivas Map e reduce. Os dados de entrada são específicos para cada aplicação, e descritos pelo usuário. A saída é um conjunto de pares no formato (chave, valor). A função Map é aplicada aos dados de entrada e produz uma lista intermediária de pares (chave, valor). Todos os valores intermediários associados a uma mesma chave são agrupados e enviados à função reduce. A função reduce é então aplicada para todos os pares intermediários com a mesma chave. A função combina esses valores para formar um conjunto menor de resultados. Tipicamente, há apenas zero ou um valores de saída em cada função reduce.

O pseudocódigo a seguir apresenta um exemplo de uso do MapReduce, cujo objetivo é contar a quantidade de ocorrências de cada palavra em um documento.

A função Map recebe como valor uma linha do documento texto, e como chave o número da linha. Para cada palavra encontrada na linha recebida, a função emite a palavra e a contagem de uma ocorrência. A função Reduce, recebe como chave uma palavra, e um iterador para todos os valores emitidos pela função Map, associados com a palavra questão. As ocorrências da palavra são agrupadas e a função retorna palavra e seu total de ocorrências.

Listing 2.1: Some Code

```
1 Function Map (Integer chave, String valor):
2   #chave: numero da linha no arquivo.
3   #valor: texto da linha correspondente.
4   listaDePalavras = split (valor)
5   for palavra in listaDePalavras:
6     emit (palavra, 1)
7 Function reduce (String chave, Iterator valores):
8   #chave: palavra emitida pela funcao Map.
9   #valores: conjunto de valores emitidos para a chave.
10  total = 0
11  for v in valores:
12    total = total + 1
13  emit (palavra, total)
```

A Figura 1 ilustra o fluxo de execução para este exemplo. A entrada é um arquivo contendo as linhas "Exemplo conta palavras"e "Hadoop exemplo palavras".

Figura 1: Fluxo simplificado da contagem de palavras com o MapReduce

Arquitetura do MapReduce

O MapReduce é constituído de uma arquitetura com dois tipos principais de nós: *Master* e *Worker*. O nó mestre tem como função atender requisições de execução dos usuários, gerenciá-las, criar tarefas e distribuí-las entre os nós trabalhadores, que executam as tarefas com base nas funções Map e Reduce definidas pelo usuário. A arquitetura também é inclui um sistema de arquivos distribuídos, onde ficam armazenados os dados de entrada e intermediários.

Visão geral do fluxo de execução

As chamadas da função Map são distribuídas automaticamente entre as diversas máquinas através do particionamento dos dados de entrada em M conjuntos. Cada conjunto pode ser processado em paralelo por diferentes máquinas. As chamadas reduce são distribuídas através do particionamento do conjunto intermediário de pares em R partes. O número de partições R pode ser definido pelo usuário.

A Figura 2 apresenta o fluxo de uma execução do MapReduce. A sequência de ações descrita a seguir explica o que ocorre em cada um dos passos. A numeração dos itens a seguir corresponde à numeração da figura.

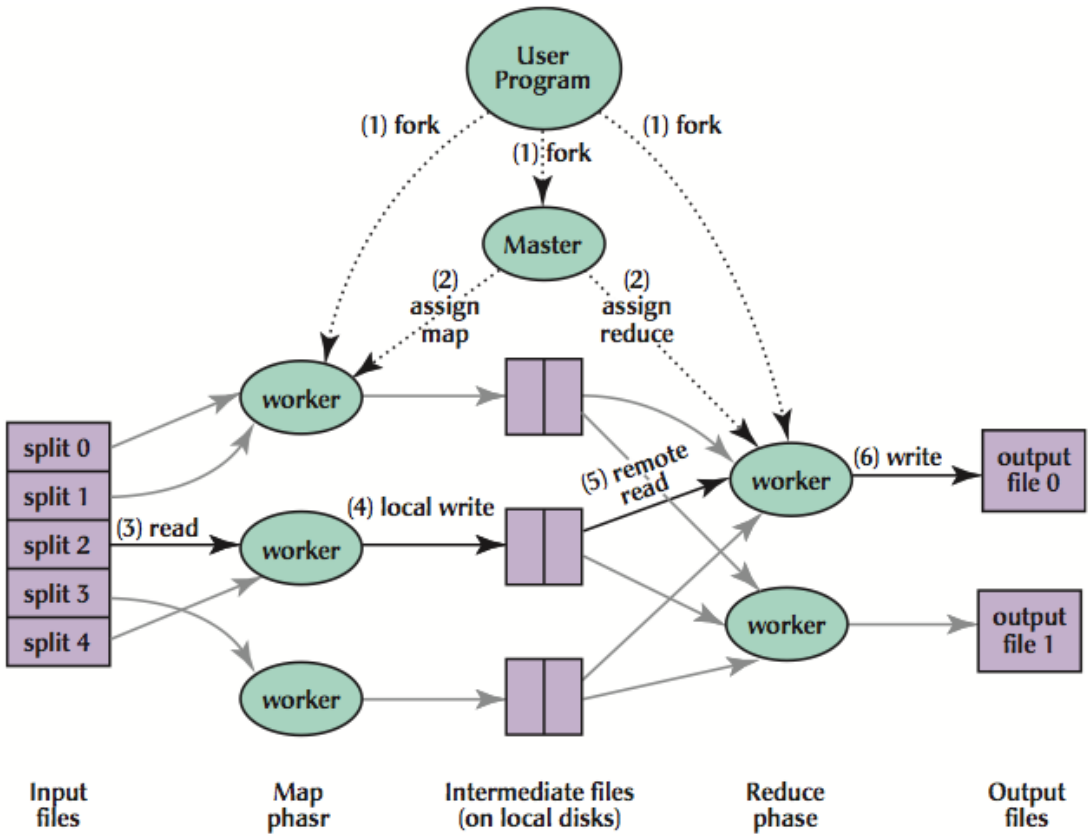

Figura 2: Visão geral do funcionamento do modelo MapReduce.

1. A biblioteca MapReduce no programa do usuário primeiro divide os arquivos de entrada em M pedaços. Em seguida, iniciam-se muitas cópias do programa em um cluster de máquinas.

2. Uma das cópias do programa é especial: o mestre (master). Os demais são trabalhadores (escravos, slaves) cujo trabalho é atribuído pelo mestre. Existem M tarefas Map e R tarefas Reduce a serem atribuídas. O mestre atribui aos trabalhadores ociosos uma tarefa Map ou uma tarefa Reduce.

3. Um trabalhador que recebe uma tarefa Map lê o conteúdo do fragmento de entrada correspondente. Ele analisa pares (chave, valor), a partir dos dados de entrada e encaminha cada par para a função Map definida pelo usuário. Os pares (chave, valor) intermediários, produzidos pela função Map, são colocados no buffer de memória;

4. Um trabalhador que recebe uma tarefa Map lê o conteúdo do fragmento de entrada correspondente. Ele analisa pares (chave, valor), a partir dos dados de entrada e encaminha cada par para a função Map definida pelo usuário. Os pares (chave, valor) intermediários, produzidos pela função Map, são colocados no buffer de memória;

5. Periodicamente, os pares colocados no buffer são gravados no disco local, divididos em regiões R pela função de particionamento. As localizações desses pares bufferizados no disco local são passadas de volta para o mestre, que é responsável pelo encaminhamento desses locais aos trabalhadores Reduce;

6. Quando um trabalhador Reduce é notificado pelo mestre sobre essas localizações, ele usa chamadas de procedimento remoto para ler os dados no buffer, a partir dos discos locais dos trabalhadores Map. Quando um trabalhador Reduce tiver lido todos os dados intermediários para sua partição, ela é ordenada pela chave intermediária para que todas as ocorrências da mesma chave sejam agrupadas. Se a quantidade de dados intermediários é muito grande para caber na memória, um tipo de ordenação externa é usado;

7. O trabalhador Reduce itera sobre os dados intermediários ordenados e, para cada chave intermediária única encontrada, passa a chave e o conjunto correspondente de valores intermediários para função Reduce do usuário. A saída da função Reduce é anexada a um arquivo de saída final para essa partição Reduce;

8. Quando todas as tarefas Map e Reduce são concluídas, o mestre acorda o programa do usuário. Neste ponto, a chamada MapReduce no programa do

usuário retorna para o código do usuário.

2.2.1 Hadoop

Uma das implementações mais conhecidas do MapReduce é o Hadoop, desenvolvido por Doug Cutting em 2005 e mantido pela Apache Software Foundation [**?**]. O Hadoop é uma implementação código aberto em Java do modelo criado pela Google, que provê o gerenciamento de computação distribuída. Sua finalidade é desenvolver software livre para computação distribuída, escalável e confiável [White 2009].

Facebook, Yahoo! e eBay utilizam o ambiente Hadoop para processar diariamente terabytes de dados e logs de eventos em seus *clusters* para detecção de spam, *business intelligence* e diferentes tipos de otimização [Cherkasova 2011].

Um dos principais benefícios do Hadoop é a sua capacidade de lidar com falhas, sejam de disco, processos, ou de nós, e permitir que o trabalho do usuário possa ser concluído.

HDFS

A plataforma Hadoop suporta diversos sistemas de arquivos distintos, como Amazon S3 (Native e Block-based), CloudStore, HAR, sistemas mantidos por servidores FTP e HTTP, Local (destinado a unidades de armazenamento conectadas localmente), mas fornece um sistema e arquivos padrão. O *Hadoop Distributed File System* (HDFS) é um sistema de arquivos distribuído desenvolvido para armazenar grandes conjuntos de dados e ser altamente tolerante a falhas [White 2009].

A arquitetura do HDFS também é do tipo mestre-escravos. O nó mestre (*NameNode*) é responsável por manter e controlar todos os metadados do sistema de arquivos e gerenciar a localização dos dados. Também é responsável por outras atividades, como por exemplo, balanceamento de carga, *garbage collection*, e atendimento a requisições dos clientes. Os nós escravos (*DataNode*) que armazenam e transmitem os dados aos usuários que os requisitarem.

A Figura 3 ilustra a arquitetura do sistema de arquivos distribuídos. O namenode gerencia e manipula todas as informações dos arquivos, tal como a loca-

lização e o acesso. Os datanodes se encarregam da leitura e escrita das informações nos sistemas de arquivos cliente.

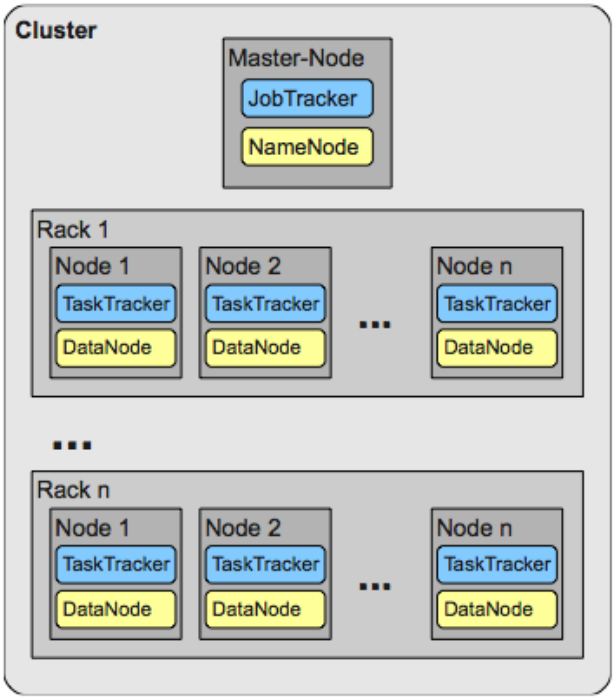

Figura 3: Visão abstrata do cluster.

O HDFS incorpora funcionalidades que têm grande impacto no desempenho geral do sistema. Uma delas é conhecida como *rack awareness*. Com esse recurso, o sistema de arquivos é capaz de identificar os DataNodes que pertencem a um mesmo *rack*, e distribuir as réplicas de maneira mais inteligente, aumentando a performance e confiabilidade do sistema. A outra é a distribuição dos dados. O sistema de arquivos busca manter um balanceamento na ocupação das unidades de armazenamento, e o *framework*, busca atribuir tarefas a um *worker* que possua, em sua unidade de armazenamento local os dados que devem ser processados. Assim, quando executa-se grandes operações MapReduce com um número significante de nós, a maioria dos dados são lidos localmente e o consumo de banda é mínimo.

2.3 Ordenação Paralela

2.1 Ordenação em memória interna A ordenação em memória interna é caracterizada pelo armazenamento de todos os registros na memória principal, onde seus acessos são feitos diretamente pelo pro- cessador. Isto é possível quando a

quantidade de dados a ser ordenada é suficientemente pequena para o processo ser realizado basicamente na memora principal. Quase todos os algoritmos de ordenação estudados pertencem a esta classificação. Tais procedimentos podem ser categorizados através de três técnicas empregadas: os métodos de inserção, seleção e partição.

2.2 Ordenação em Memória Externa

Quando o número de registros a ser ordenado é maior do que o computador pode suportar em sua memória principal, surgem problemas interessantes a serem estuda- dos. Embora o fato de a ordenação em memória interna e externa possuírem o mesmo objetivo, elas são muito diferentes, uma vez que o acesso eficiente aos arquivos muito extensos é bastante limitado. As estruturas de dados devem ser organizadas de modo que lentos dispositivos periféricos de memória, geralmente discos rígidos, possam rapidamente lidar com as exigências do algoritmo. Consequentemente, a maioria dos algoritmos de ordenação interna é praticamente inútil para a ordenação externa. Por conseguinte, se torna necessário repensar toda a questão[7].

A ordenação externa deve ser aplicada em conjuntos de dados compostos por um nú- mero de registros maior do que a memória interna do computador pode armazenar. Os métodos de ordenação externa são muito diferentes dos métodos de ordenação interna, apesar de, em ambos os casos, o problema ser o mesmo: rearranjar os registros de um arquivo em ordem ascendente ou descendente. Isso ocorre, pois na ordenação externa as estruturas de dados têm que levar em conta o fato de que os dados estão armazena- dos em unidades de memória externa, tais como fitas e discos magnéticos, relativamente mais lentas do que a memória principal em dois aspectos: velocidade de acesso e tipo de acesso [Ziviani 2007, Knuth 1973]. Os dados são armazenados na memória externa como um arquivo sequencial, em que apenas um registro pode ser acessado em um dado momento. Essa forte restrição torna os métodos de ordenação interna inadequados para a ordenação externa. Surge então, a necessidade de se utilizar técnicas de ordenação completamente diferentes [Ziviani 2007, Knuth 1973]. Na ordenação externa, os itens que não estão na memória principal devem ser buscados em memória secundária e trazidos para a memória principal, para assim serem comparados. Esse processo se repete inúmeras vezes, o que o torna lento, uma vez que os processadores ficam grande parte do tempo ociosos à espera da chegada dos dados à memória principal para serem processados. Por esse motivo, a grande ênfase de um método de ordenação externa deve ser na

minimização do número de vezes que cada item é transferido entre a memória interna e a memória externa. Além disso, cada transferência deve ser realizada de forma tão eficiente quanto as características dos equipamentos disponíveis permitam [Ziviani 2007].

2.4 Algoritmos de ordenação Paralela

Condições de implementação de algoritmos paralelos de ordenação

• **Habilidade de explorar distribuições iniciais parcialmente ordenadas** Alguns algoritmos podem se beneficiar de cenários nos quais a sequência de entrada dos dados é mesma, ou pouco alterada. Nesse caso, é possível obter melhor desempenho ao realizar menos trabalho e movimentação de dados. Se a alteração na posição dos elementos da sequência é pequena o suficiente, grande parte dos processadores mantém seus dados iniciais e precisa se comunicar apenas com os processadores vizinhos.

Movimentação dos dados A movimentação de dados entre processadores deve ser mínima durante a execução do algoritmo. Em um sistema de memória distribuída, a quantidade de dados a ser movimentada é um ponto crítico, pois o custo de troca de dados pode dominar o custo de execução total e limitar a escalabilidade.

Balanceamento de carga O algoritmo de ordenação paralela deve assegurar o balanceamento de carga ao distribuir os dados entre os processadores. Cada processador deve receber uma parcela equilibrada dos dados para ordenar, uma vez que o tempo de execução da aplicação é tipicamente limitada pela execução do processador mais sobrecarregado.

Latência de comunicação A latência de comunicação é definida como o tempo médio necessário para enviar uma mensagem de um processador a outro. Em grandes sistemas distribuídos, reduzir o tempo de latência se torna muito importante.

Sobreposição de comunicação e computação Em qualquer aplicação paralela, existem tarefas com focos em computação e comunicação. A sobreposição de

tais tarefas permite que sejam feitas tarefas de processamento e ao mesmo tempo operações de entrada e saída de dados, evitando que os recursos fiquem ociosos durante o intervalo de tempo necessário para a transmissão da carga de trabalho.

2.4.1 Sample Sort

O algoritmo de Ordenação por Amostragem [White 2009, Venner 2009] é um método de ordenação externa para ordenar grandes quantidades de dados. A ideia desse algoritmo é dividir um conjunto de dados a ser ordenado em subconjuntos (partições) de forma que os elementos da partição à esquerda sejam menores do que os elementos à direita da partição, isto é, todas as chaves da partição i são menores que as chaves da partição i + 1. Cada uma das partições é designada a um processador para ordenação (ordenação interna) e, assim, é produzido um conjunto de arquivos ordenados que, se concatenados, formam um arquivo globalmente ordenado.

A divisão dos dados em partições (particionamento) é feita por meio de uma estratégia de amostragem. Essa estratégia baseia-se na análise de um subconjunto de dados, denomi- nado amostra, ao invés de todo o conjunto, para estimar a distribuição de chaves e, assim, construir partições balanceadas.

Existem três tipos de estratégias de amostragem: SplitSampler, IntervalSampler e Ran- domSampler. O SplitSampler seleciona apenas os n primeiros registros do arquivo para for- mar a amostra. Já o IntervalSampler forma a amostra por meio da escolha de chaves em intervalos regulares no arquivo. No RandomSampler, a amostra é constituída por chaves selecionadas aleatoriamente no conjunto.

Após a formação das amostras, são definidos os limites para as partições, isto é, o intervalo de valores compreendido por cada partição. A quantidade de limites amostrados (qla) é determinada pelo equação: qla = np - 1, onde np é o número de partições (np = nc x nmaq, sendo nc o número de cores e nmaq o número de máquinas). As informações das partições são armazenadas em um arquivo e transmitidas para as máquinas participantes por meio de cache distribuído.

O melhor tipo de estratégia de amostragem é determinado pela entrada de dados. Para arquivos quase ordenados, o SplitSampler não é recomendado, uma vez que não seleciona as chaves de todo o conjunto e, assim, não produz bons resultados.

Nesse caso, a melhor escolha é o IntervalSampler pelo fato de selecionar chaves que representam melhor a distribuição do conjunto. O RandomSampler é considerado um bom amostrador de propósito geral [White 2009], por isso será utilizado neste trabalho.

O RandomSampler apresenta os seguintes parâmetros de balanceamento: probabilidade de escolha de uma chave, o número máximo de amostras a serem selecionadas para realizar a amostragem e o número máximo de partições que podem ser utilizadas. Ele termina quando algum dos parâmetros é atingido.

A Figura 4.6 apresenta um esquema da estrutura de funcionamento do algoritmo, quando implementado em MapReduce no ambiente Hadoop. O algoritmo pode ser di- vidido em duas fases: Map e Reduce.

Na fase Map, os arquivos de entrada são lidos, ocorre a formação do vetor inicial e dos pares (chave, valor) dos valores lidos (passos 1.1 e 1.2). Depois disso, os dados são divididos em partições cujo número é determinado pelo número de máquinas e seus núcleos de processamento (passo 1.3). No exemplo apresentado, são utilizados dois núcleos de pro- cessamento (cores). Portanto, existem duas partições e, dessa forma, deve ser escolhido apenas um limite para as partições. Esse limite é definido por meio da seleção do ele- mento que melhor representa a distribuição de chaves na amostra formada pela estratégia RandomSampler. Para esse exemplo o limite escolhido foi o número 13, o que implica que em uma partição estarão valores menores e iguais a 13 e, na outra, valores maiores que 13. Por meio de cache distribuído, as informações das partições são transmitidas para as máquinas participantes e os dados particionados. Cada partição é então atribuída a uma tarefa Reduce (processador) diferente.

Na fase Reduce, os dados são ordenados localmente (passo 2.1), ou seja, em cada processador os dados são ordenados na memória interna. Para essa ordenação, avalia-se a profundidade da árvore de recursão e, se ela for baixa, utiliza-se o algoritmo QuickSort. Caso contrário, o HeapSort é utilizado [White 2009]. Após isso, os dados retornam para a máquina master, onde são concatenados, formando o vetor ordenado.O balanceamento das partições, ou seja, a formação de partições aproximadamente iguais no tamanho é muito importante para o algoritmo de Ordenação por Amostragem, pois evita que os tempos de ordenação sejam dominados por um único processador. Em outras palavras, o equilíbrio das partições reduz a possibilidade de

que um processador esteja ocioso, enquanto outro processador está sobrecarregado, situação que comprometeria o desempenho do algoritmo [White 2009].

2.4.2 Quick Sort

3 Desenvolvimento

3.1 Metodologia

O início do projeto será destinado ao estudo mais detalhado da computação paralela, em especial os algoritmos de ordenação paralela, dos fatores que influenciam o desempenho de tais algoritmos, o modelo MapReduce e a plataforma Hadoop. O passo seguinte é conhecer detalhadamente o algoritmo paralelo a ser implementado e definir as estratégias para sua implentação ambiente Hadoop. O algoritmo implementado deve ser cuidadosamente avaliado para verificar um funcionamento adequado com diferentes entradas e número de máquinas.

Em seguida, serão realizados experimentos para testes de desempenho dos algoritmos com relação à quantidade de máquinas, quantidade de dados e conjunto de dados. Os resultados obtidos serão analisados e permitirão comparar a desempenho dos algoritmos em cada situação.

3.2 Infraestrutura

A infra estrutura necessária ao desenvolvimento do projeto será fornecida pelo Laboratório de Redes e Sistemas (LABORES) do Departamento de Computação (DECOM). O laboratório possui um *cluster* formado por cinco máquinas Dell Optiplex 380, que serão utilizadas na realização dos testes dos algoritmos. Os algoritmos serão desenvolvidos em linguagem Java, de acordo com o modelo MapReduce, no ambiente Hadoop.

Cada máquina do *cluster* apresenta as seguintes características:

- Processador Intel Core 2 Duo de 3.0 GHz
- Disco rígido SATA de 500 GB 7200 RPM
- Memória RAM de 4 GB
- Placa de rede Gigabit Ethernet

- Sistema operacional Linux Ubunbu 10.04 32 bits
- Sun Java JDK 1.6.0 19.0-b09
- Apache Hadoop 1.0.2

3.3 Descrição dos experimentos

A primeira parte dos experimentos consitiu em reproduzir os resultados já encontrados no trabalho de referência: testes de ordenação com os *benchmarks* TeraSort e Sort, e com o algoritmo Ordenação por Amostragem. Em todos os casos, os testes são compostos de duas partes: geração da carga de dados, seguida da ordenação. A seguir são discutidos os testes.

3.3.1 Testes com benchmarks: TeraSort e Sort

Os *benchmarks* TeraSort e Sort foram os primeiros testes de ordenação realizados. O uso de algoritmos conhecidos e consolidados na ordenação no ambiente Hadoop permite compreender o funcionamento dos algoritmos e do ambiente dos testes.

Terasort

This package consists of 3 map/reduce applications for Hadoop to compete in the annual terabyte sort competition.

* TeraGen is a map/reduce program to generate the data. * TeraSort samples the input data and uses map/reduce to sort the data into a total order. * TeraValidate is a map/reduce program that validates the output is sorted.

O TeraSort consiste de três algoritmos, que são responsáveis pela geração dos dados, ordenação e validação.

A geração dos dados é feita pelo algoritmo TeraGen. Os registros gerados têm um formato específico, descrito na Figura **??**. O registro é formado por uma chave, um id e um valor.

Chave as chaves são caracteres aleatórios do conjunto ’ ’ .. ’ ’.

Id um valor inteiro

Valor consiste de 70 caracteres de 'A' a 'Z'.

O número de registros gerados é um parâmetro definido pelo usuário, e os dados gerados são divididos em dois arquivos. Nos testes realizados, foram gerados dois arquivos, cada um contendo 50 mil linhas.

O TeraSort lê tais arquivos e realiza a ordenação. Após a ordenação, os dados são validados pelo TeraValidade. Caso haja algum erro na ordenação, o algoritmo escreve um arquivo informando quais foram as chaves com erros.

Sort

Os dados utilizados para ordenação com o Sort foram gerados pelo algoritmo RandomWriter. Para cada máquina do *cluster*, são escritos 10 arquivos de 1GB cada em formato binário, totalizando 10GB.

3.3.2 Testes com o Algoritmo Ordenação por Amostragem

(GeraDados) Programa implementado em Java para geração de chaves inteiras aleatórias. Foram gerados 10 conjuntos de chaves entre 10_6 (2MB) e 10_{10} (20GB) chaves inteiras. Anotar os tempos de execução de cada algoritmo.

Variando o conjunto de dados

(4 máquinas)

Objetivo: avaliar a influência dos valores gerados aleatoriamente no desempenho do algoritmo. Testes com 10 conjuntos de 10_6 dados. Para cada conjunto, executar 10 vezes com os parâmetros de balanceamento descritos anteriormente.

Variando a quantidade de dados

(4 máquinas)

Objetivo: avaliar a complexidade do algoritmo quando o conjunto de dados a serem ordenados aumenta. Testes com dados de 10_6 a 10_{10} gerados aleatoriamente.

Para cada quantidade de dados, o algoritmo foi executado três vezes com os parâmetros descritos anteriormente.

Variando a quantidade de máquinas

(2 a 5 máquinas)

Objetivo: avaliar a escalabilidade do algoritmo (diminuição do tempo de ordenação) Testes com o mesmo conjunto de 10_8 dados em diferentes quantidades de máquinas, de 2 a 5.

Para cada quantidade de máquinas, o algoritmo foi executado três vezes, com os parâmetros de balanceamento descritos anteriormente.

Parâmetros do algortimo

Frequência: Número max de amostras: 10 mil Núm max de partições: 4 (5 máquinas) 6 (5 máquinas) 8 (5 máquinas) 10 (5 máquinas)

3.4 Cronograma de trabalho

O cronograma de trabalho inclui as atividades que devem ser realizadas e como elas devem ser alocadas durante as disciplinas TCC I e TCC II para que o projeto possa ser concluído com sucesso. As tarefas a serem desenvolvidas estão descritas a seguir:

1. Pesquisa bibliográfica sobre o tema do projeto e escrita da proposta
2. Estudo mais detalhado dos algoritmos de ordenação paralela, modelo MapReduce e Hadoop.
3. Configuração do ambiente Hadoop no laboratório.
4. Implementação e testes.
5. Escrita, revisão e entrega do relatório.
6. Análise comparativa entre os resultados.
7. Escrita e revisão do projeto final.
8. Entrega e apresentação.

Na Tabela 1 está descrito o cronograma esperado para o desenvolvimento do projeto. Cada atividade foi alocada para se adequar da melhor maneira ao tempo disponível, mas é possível que o cronograma seja refinado posteriormente, com a inclusão de novas atividades ou redistribuição das tarefas existentes.

Atividade	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov
1	•	•								
2		•	•							
3			•							
4			•	•		•	•			
5				•	•					
6								•		
7									•	
8										•

Tabela 1: Cronograma proposto para o projeto

4 Resultados Preliminares

5 Conclusões e Propostas de Continuidade

Referências Bibliográficas

[Aho et al. 1974]Aho, A. V.; Hopcroft, J. E.; Ullman, J. D. *The Design and Analysis of Computer Algorithms.* : Addison-Wesley, 1974.

[Amato et al. 1998]Amato, N. M.; Iyer, R.; Sundaresan, S.; Wu, Y. *A Comparison of Parallel Sorting Algorithms on Different Architectures.* College Station, TX, USA, 1998.

[Asanovic et al. 2009]Asanovic, K.; Bodik, R.; Demmel, J.; Keaveny, T.; Keutzer, K.; Kubiatowicz, J.; Morgan, N.; Patterson, D.; Sen, K.; Wawrzynek, J.; Wessel, D.; Yelick, K. A view of the parallel computing landscape. *Commun. ACM*, ACM, New York, NY, USA, v. 52, n. 10, p. 56–67, out. 2009.

[Bryant 2011]Bryant, R. E. Data-Intensive Scalable Computing for Scientific Applications. *Computing in Science and Engineering*, IEEE Computer Society, Los Alamitos, CA, USA, v. 99, n. PrePrints, 2011.

[Cherkasova 2011]Cherkasova, L. Performance modeling in mapreduce environments: challenges and opportunities. In: *Proceedings of the second joint WOSP/SIPEW international conference on Performance engineering.* New York, NY, USA: ACM, 2011. (ICPE '11), p. 5–6.

[Dean e Ghemawat 2008]Dean, J.; Ghemawat, S. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, ACM, New York, NY, USA, v. 51, n. 1, p. 107–113, jan. 2008.

[Ernst et al. 2009]Ernst, D.; Wittman, B.; Harvey, B.; Murphy, T.; Wrinn, M. Preparing students for ubiquitous parallelism. In: *Proceedings of the 40th ACM technical symposium on Computer science education.* New York, NY, USA: ACM, 2009. (SIGCSE '09), p. 136–137.

[Gantz 2008]Gantz, J. *The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011.* 2008.

[Kale e Solomonik 2010]Kale, V.; Solomonik, E. Parallel sorting pattern. In: *Proceedings of the 2010 Workshop on Parallel Programming Patterns*. New York, NY, USA: ACM, 2010. (ParaPLoP '10), p. 10:1–10:12.

[Lin e Dyer 2010]Lin, J.; Dyer, C. *Data-Intensive Text Processing with MapReduce*. : Morgan & Claypool Publishers, 2010. (Synthesis Lectures on Human Language Technologies).

[Manferdelli et al. 2008]Manferdelli, J. L.; Govindaraju, N. K.; Crall, C. Challenges and opportunities in Many-Core computing. *Proceedings of the IEEE*, , v. 96, n. 5, p. 808–815, may 2008.

[Pinhão 2011]Pinhão, P. de M. *Ordenação Paralela no Ambiente Hadoop*. 2011.

[Prinslow 2011]Prinslow, G. *Overview of Performance Measurement and Analytical Modeling Techniques for Multi-core Processors*. 2011.

[Ranger et al. 2007]Ranger, C.; Raghuraman, R.; Penmetsa, A.; Bradski, G.; Kozyrakis, C. Evaluating mapreduce for multi-core and multiprocessor systems. In: *Proceedings of the 2007 IEEE 13th International Symposium on High Performance Computer Architecture*. Washington, DC, USA: IEEE Computer Society, 2007. (HPCA '07), p. 13–24.

[Satish et al. 2009]Satish, N.; Harris, M.; Garland, M. Designing efficient sorting algorithms for manycore gpus. In: *Proceedings of the 2009 IEEE International Symposium on Parallel&Distributed Processing*. Washington, DC, USA: IEEE Computer Society, 2009. p. 1–10.

[White 2009]White, T. *Hadoop: The Definitive Guide*. first edition. : O'Reilly, 2009.