MapReduce

Computação Distribuída Intensiva em Dados

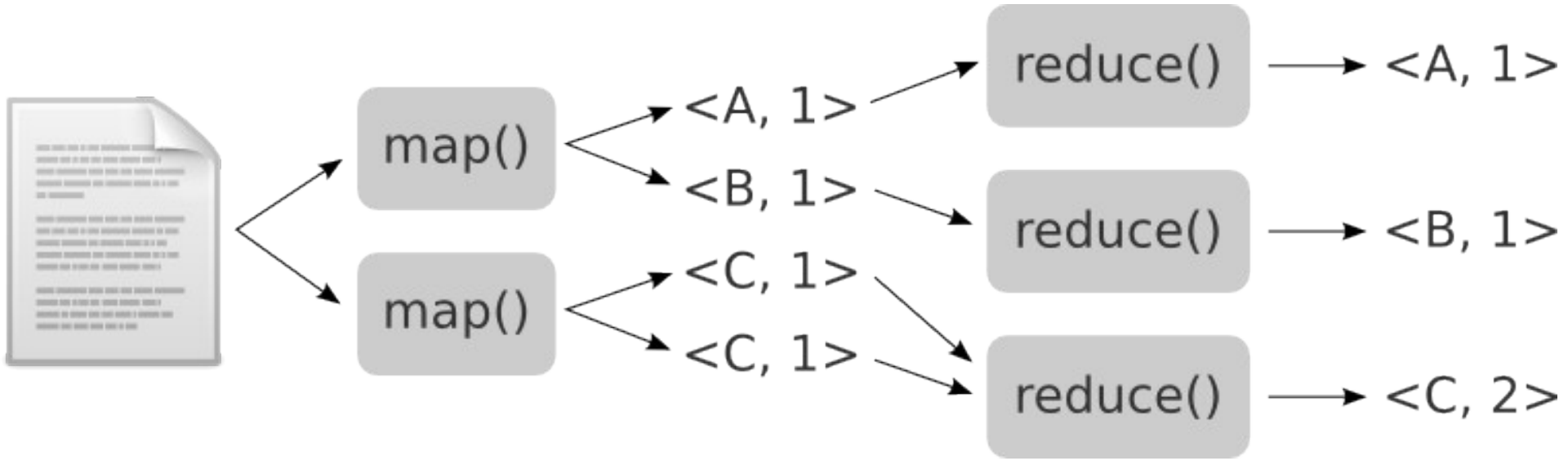
O que é o MapReduce?

- Um **modelo de programação** criado pela Google.
- Voltado ao processamento de **grandes volumes de dados**.

Qual o modelo a seguir?

- Aplicativos desenvolvidos na Google seguiam um **comportamento semelhante**.
  - Segmentos dos dados de entrada eram associados a **chaves**.
  - Algum processamento era realizado sobre os **valores/segmentos** associados a uma mesma chave.
- Este comportamento era refeito para cada aplicação.

O Modelo MapReduce



Pseudocódigo

```
function map(line, text):
    foreach word in text:
        emit(word, 1)


function reduce(word, values[]):
    sum = 0
    foreach v in values:
        sum += v
```

Ex. 1: PageRank - Referências

- Map()
  - Recebe uma página Web.
  - Procura por *links*.
  - Para cada *link* emite <domínio, 1>.
- Reduce()
  - Soma todos os valores “1” de um domínio.
  - Emite <domínio, total>.

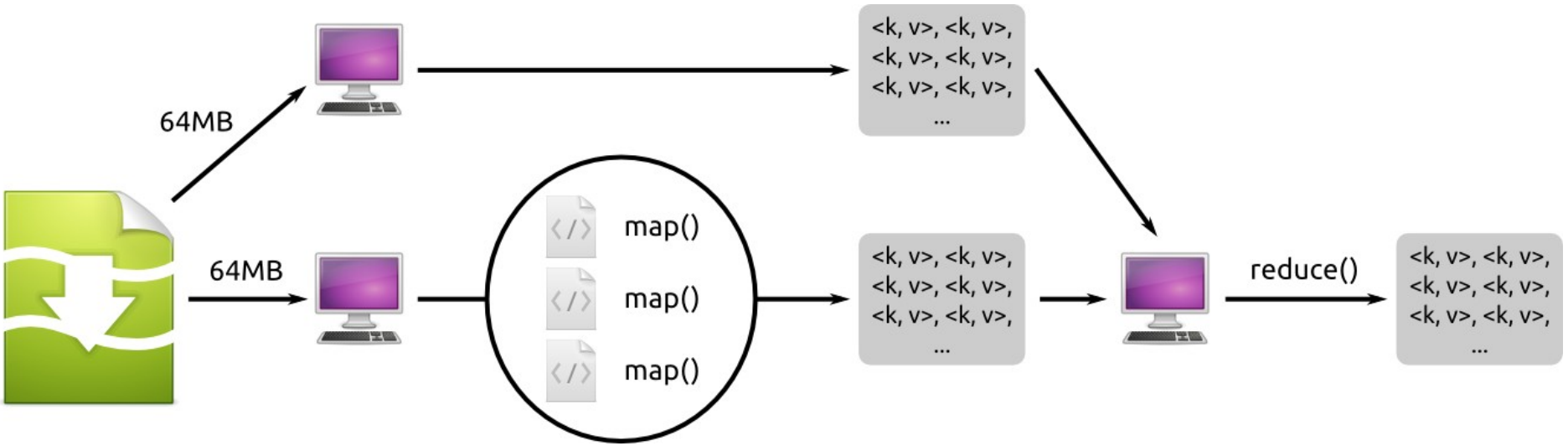
Ex. 2: Processamento de Log

- **Caso de uso:** Loja Online – média de gastos de cada usuário para um tipo de produto.
- Map()
  - Recebe *logs* de compras de vários usuários.
  - Verifica o tipo do produto (filtro).
  - Emite <usuário, valor>.
- Reduce()
  - Calcula a média dos gastos.
  - Emite <usuário, média>.

Framework MapReduce

- Os dados de entrada são armazenados em um **Sistema de Arquivos Distribuído (DFS)**.
- Os nós do DFS são as mesmas máquinas que processam os aplicativos MapReduce.
- Tarefas são escalonadas considerando a localidade dos *chunks* (“blocos”) da entrada.

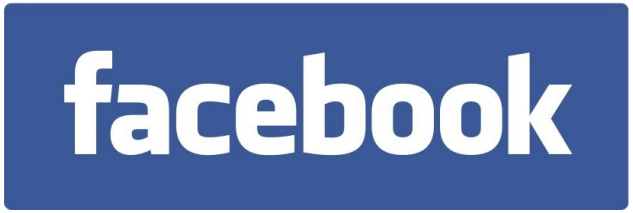
Ambiente Distribuído



Implementações

- O **Hadoop** é a implementação *open source* mais utilizada no mercado.
  - Mantido pela Apache Software Foundation.
  - Hadoop Distributed File System.
  - Desenvolvido em Java.
  - Hadoop Streaming.
- Outras implementações: GPU, Cloud, ...

Alguns Usuários...



http://wiki.apache.org/hadoop/PoweredBy

Grupo MapReduce

- 3 alunos de mestrado e 1 bolsista de IC.
- Projetos e trabalhos:
  - Adaptação do MR para Desktop Grids;
  - Simulador MRSG;
  - MR Cloud.