PrIntCloud

Disciplina: Procedência de Dados e Data Warehousing.

Aluna: Shermila Guerra Santa Cruz.

16/04/13

Roteiro

1. Fundamentação Teórica
   - A.-Cloud Computing
   - B.-Hadoop
   - C.-MapReduce
   - D.-NoSql
2. Proposta do projeto PrIntCloud.
3. Desafios.

A.- Cloud Computing

Existem servidores com níveis de uso de capacidade computacional bastantes baixos, com medias 5-10%, em períodos de pico usa o 30% a 40%. Se observa uma significativa ociosidade dos ciclos de CPU.

A.- Arquitetura Cloud Computing

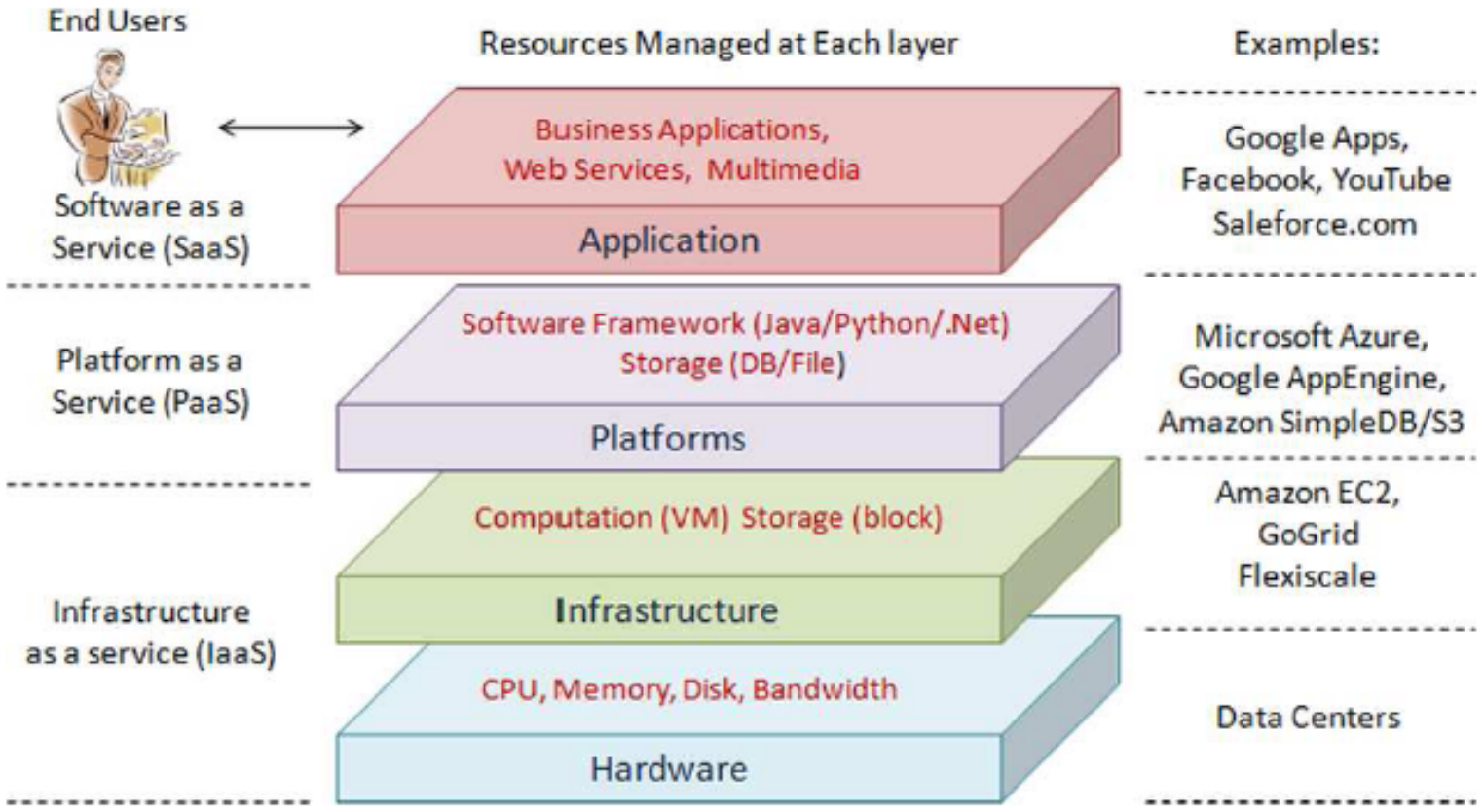


Figura extraída de Cloud computing: state-of-the-art and research challenges
Springer Journal of Internet Services and Applications, April 2010.

Gerenciamento de dados na Nuvem

Escalabilidade.

Elasticidade.

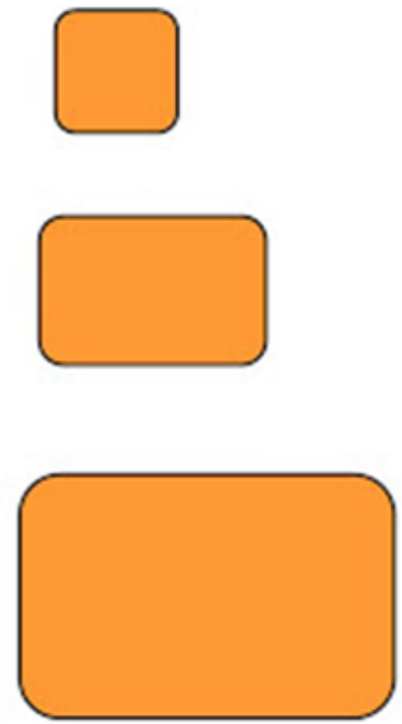
Disponibilidade.

A.-O que é escalabilidade?

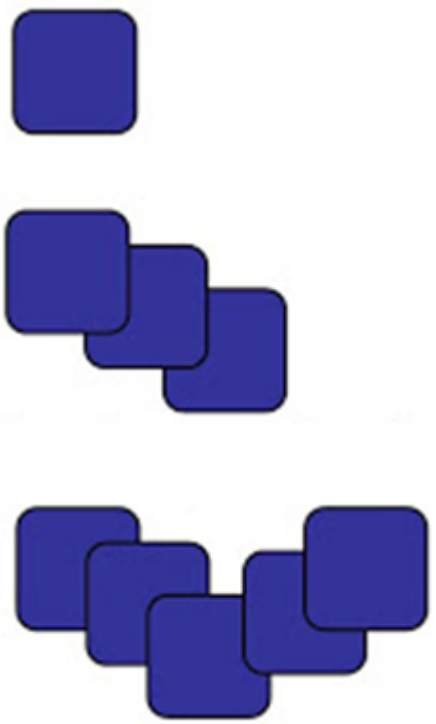
Se tem duas opções :

1. Escalabilidade Vertical.
2. Escalabilidade Horizontal.

A.- Escalabilidade Vertical Versus Horizontal.

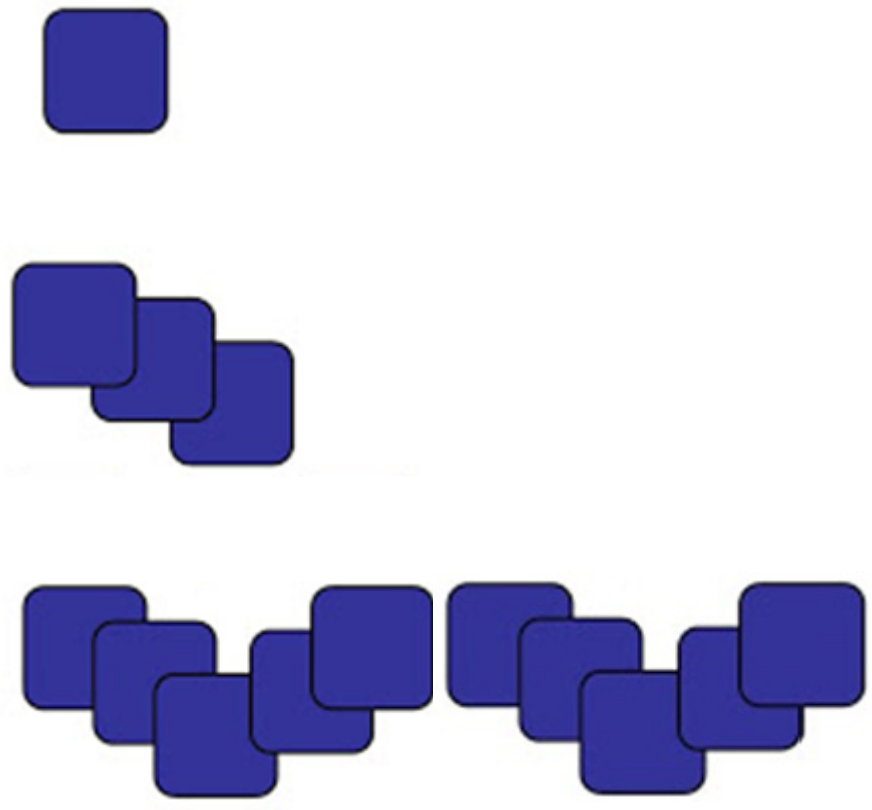


Escalabilidade Vertical



Escalabilidade Horizontal

A.- Elasticidade e disponibilidade



Motivação

Computação paralela não é trivial

- Rede comum
- Escalonamento das subtarefas
- Balanceamento de carga

Apache Hadoop

- Retira a complexidade na computação de alto desempenho
- Máquinas comuns

B.-O que é o Hadoop?

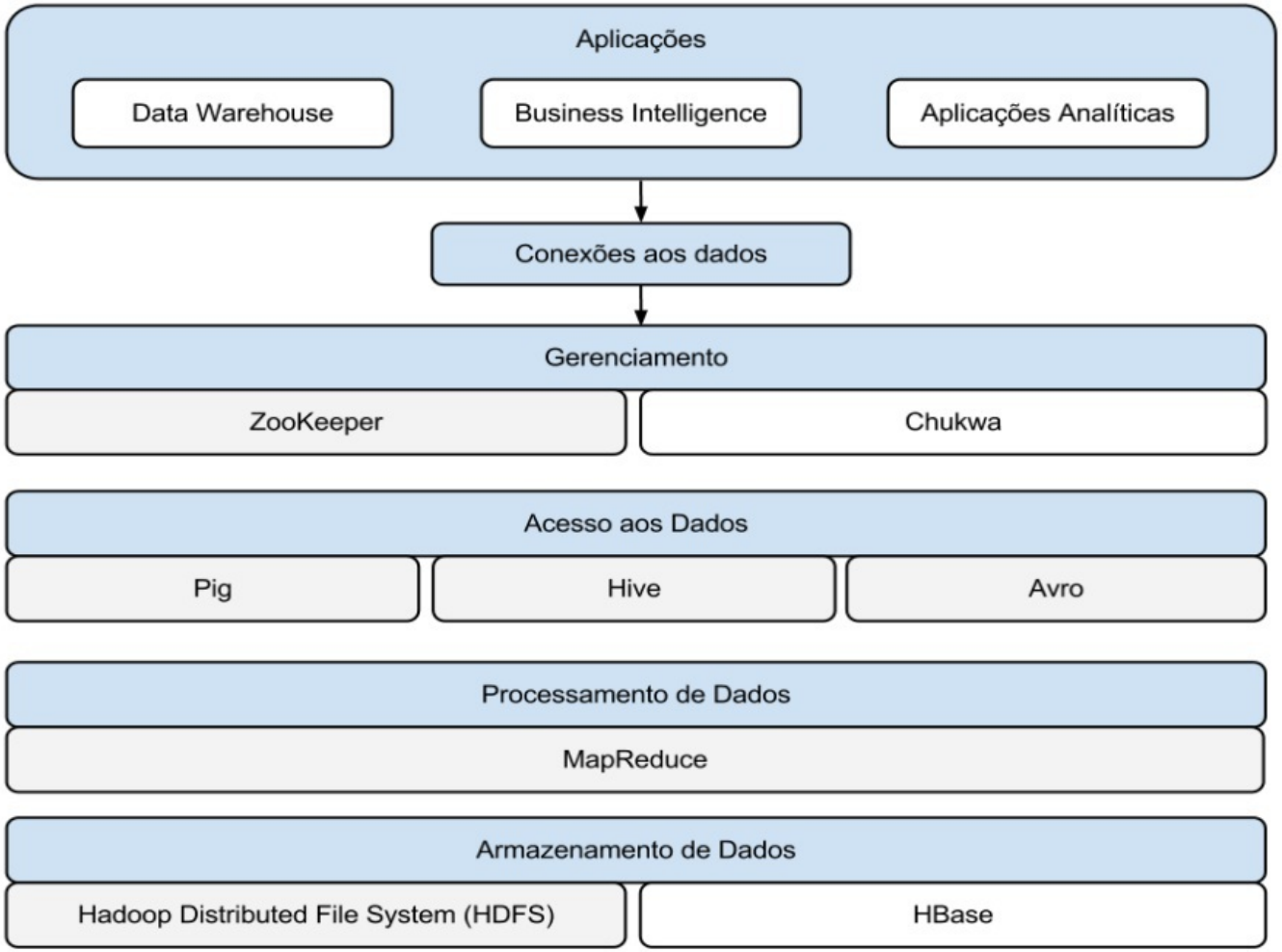
Arcabouço para processamento e armazenamento de dados em larga escala:

- Código aberto
- Implementado em Java
- Inspirado no GFS e MapReduce da Google
- Projeto principal da Fundação Apache
- Tecnologia recente, porém já muito utilizada

b.- Onde utilizar o Hadoop

- DataWarehouse
- Business Intelligence
- Aplicações analíticas
- Mídias sociais

Subprojetos do Hadoop (Rogers,2011)

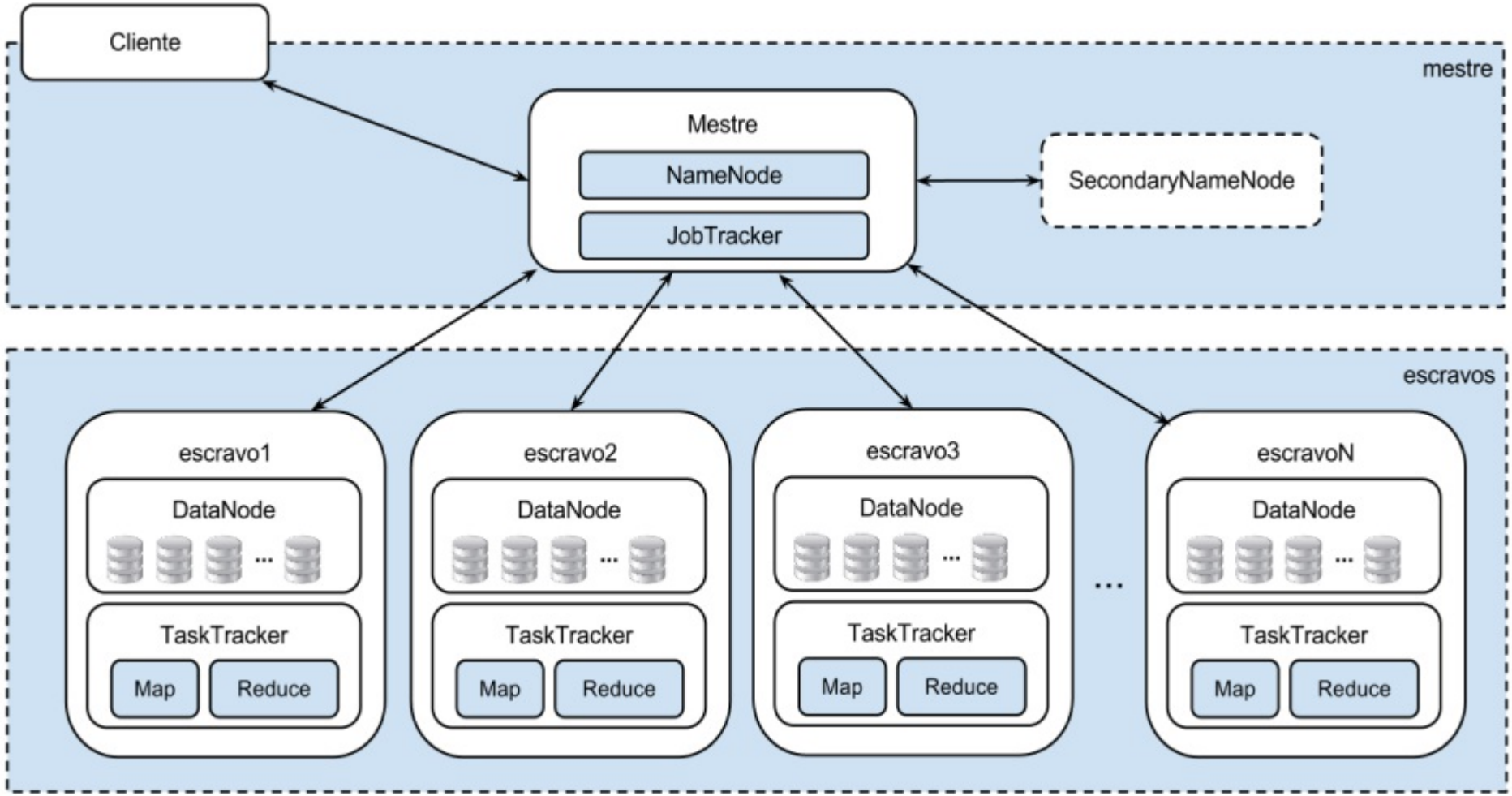


B.-Componentes do Hadoop

- Nó Mestre:
- NameNode
- DataNode
- SecondaryNameNode

- Nó(s) Escravo(s):
- JobTracker
- TaskTracker

B.-Processos do Hadoop



B.-HDFS

- Hadoop Distributed Filesystem
- Características
- Divisão em blocos
- Replicação de dados

B.-Características HDFS

- Sistema de arquivos distribuídos
- Arquitetura Mestre/Escravo
- Inspirado no Google FileSystem (GFS)
- Implementado em Java
- Armazenamento de grandes volumes de dados
- Recuperação de dados transparente ao usuário

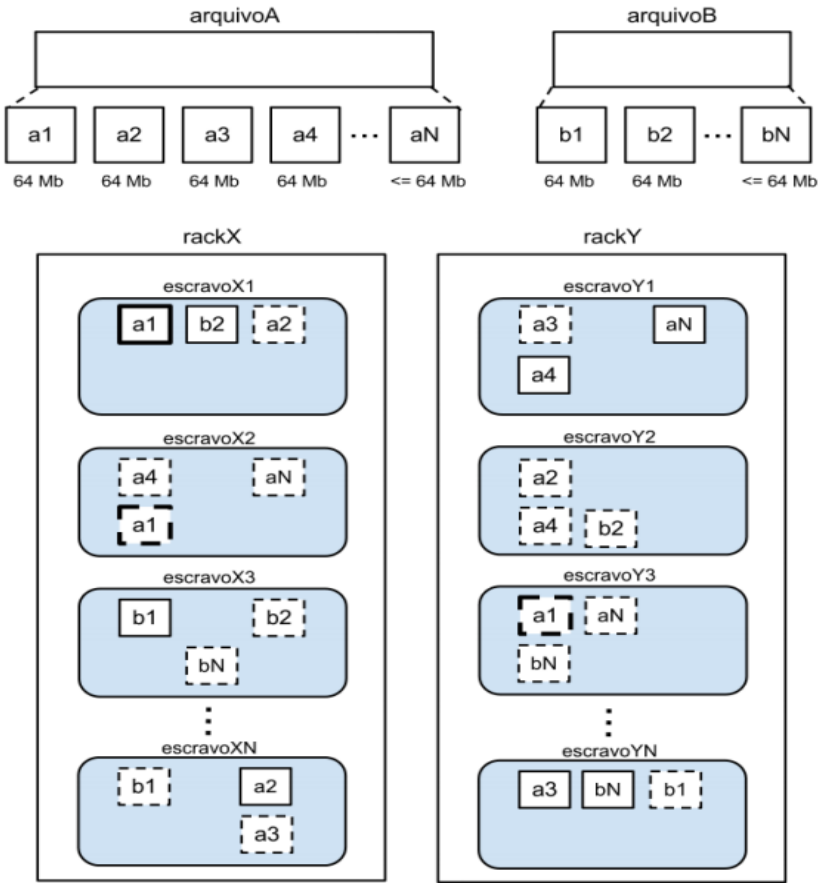
B.-Replicação de bloco de Dados

Três réplicas para cada bloco

- Aumento de segurança e disponibilidade
- Cada réplica em um diferente nó
- Dois em um mesmo armário (rack) e 1 em um armário diferente

- Re-Replicação
- Em casos de corromper uma das réplicas

B.-Replicação de bloco de Dados



MapReduce

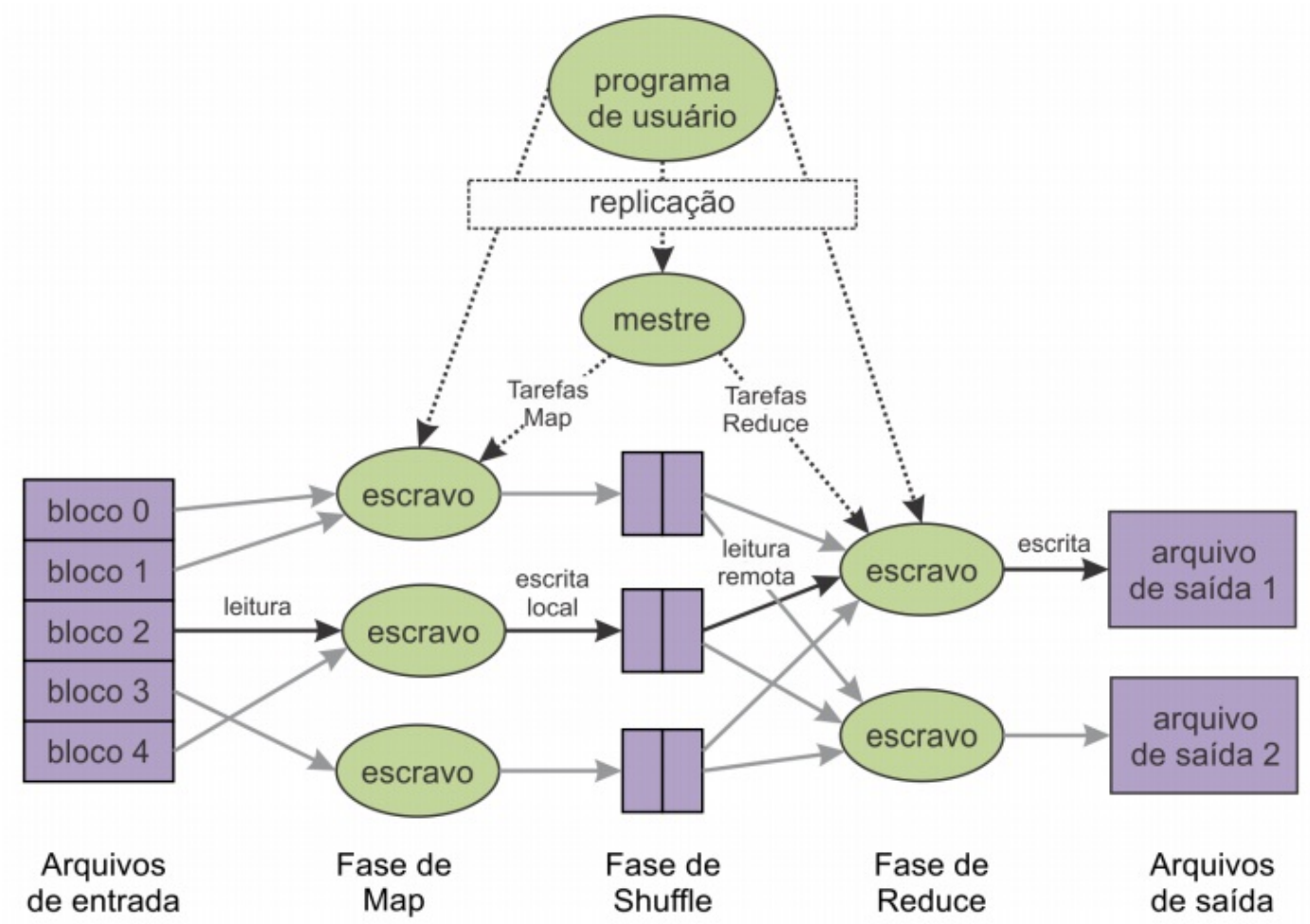
- No Hadoop é a parte do arcabouço responsável pelo processamento distribuído (paralelo) de grandes conjuntos de dados.
- O paradigma MapReduce é adequado para trabalhar com grandes quantidades de dados
- Realiza computação sobre os dados (pouca movimentação de dados)
- Utiliza os blocos armazenados no DFS, logo não necessita divisão dos dados

MapReduce

Como resolver um problema com MapReduce?

- Leia uma grande quantidade de dados
- Aplique a função **MAP**: extrai alguma informação de valor!
- Fase intermediária: Shuffle & Sort
- Aplique a função **REDUCE**: reúne, compila,
- filtra, transforma,...
- Grava os resultados

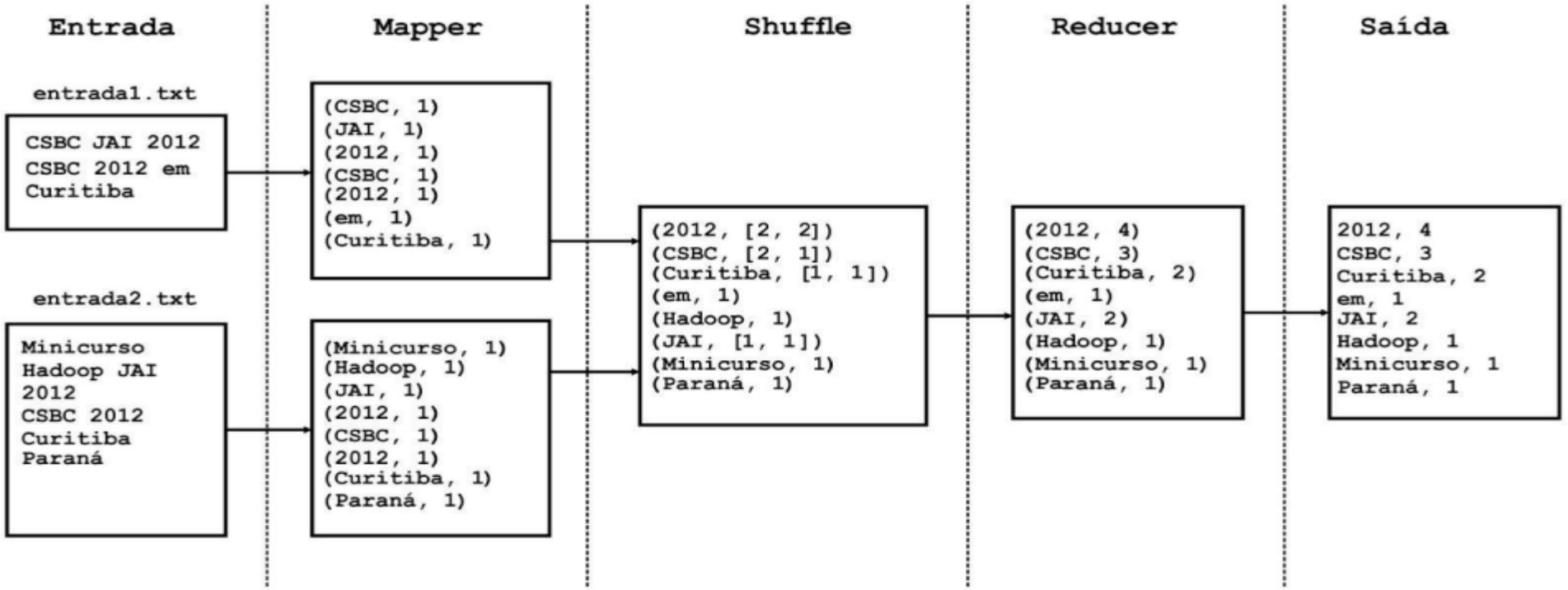
MapReduce implementado pela Google



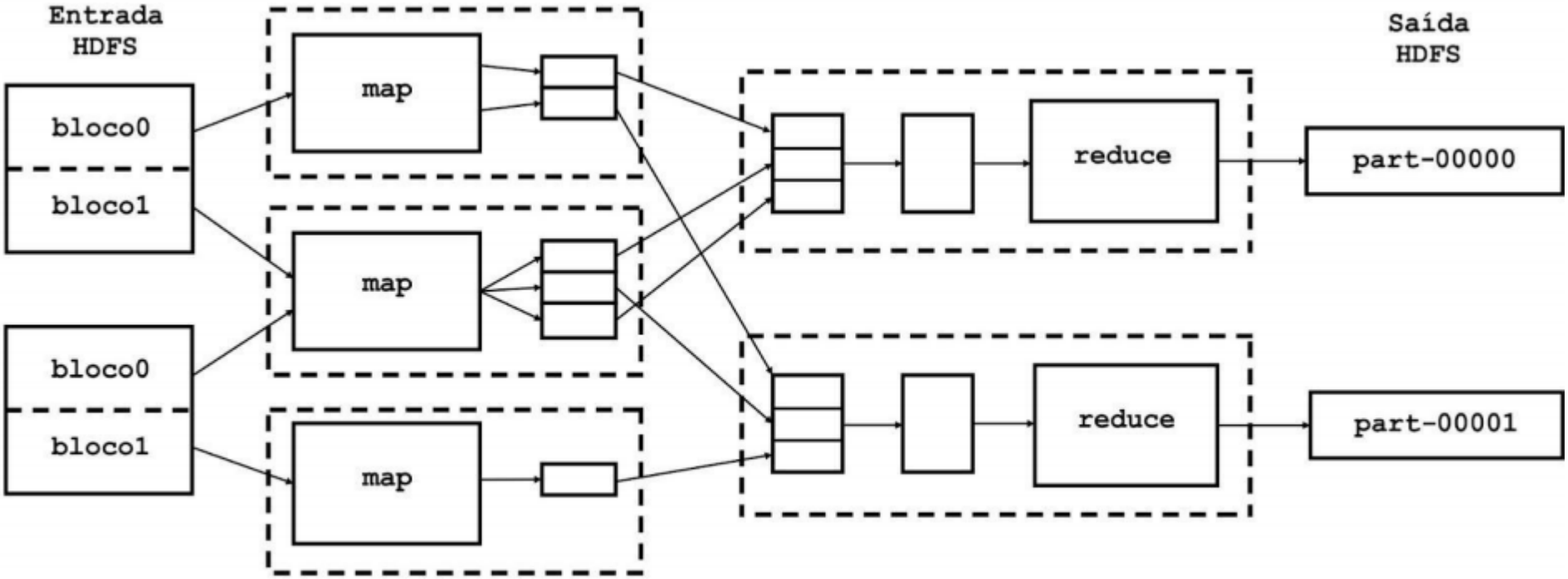
Exemplos: Word Count

- Lê arquivos texto e conta a frequência das palavras
- Entrada: arquivos texto
- Saída: arquivo texto
- Cada linha: palavra, separador (tab), quantidade
- Map: gera pares de (palavra, quantidade)
- Reduce: para cada palavra, soma as quantidades

Fluxo lógico de execução da aplicação Map Reduce Word Count



Fluxo de execução de uma aplicação MapReduce no Hadoop



D.-NoSql

D.-.-Caraterísticas do NoSql

Esquema livre.

Distribuídos .

Escalabilidade horizontal.

Consistência eventual(Not ACID).

Fácil suporte replicação .

Fonte: http://nosql-database.org/
based on 5 sources, 11 constructive feedback emails and 1 disliking comment.

2.-PrIntCloud.

Justificação

Devido à **natureza dinâmica** e a necessidade de **alto poder de processamento computacional** requerida em processos de integração, propõe-se neste projeto de pesquisa **adaptar e estender** o modelo PrInt para integração de dados em ambientes em nuvem.

Problemática

- Na **etapa de Coleta dos Dados** não proporciona escalabilidade em termos do Número de fontes de dados.

- Na **etapa de integração de dados** não propicia a escalabilidade em termos do número de operações que executa para resolver as inconsistências destas fontes de dados.

Objetivo 1

Adaptar e estender o modelo PrInt para permitir a integração de dados em nível de instância em um **ambiente na nuvem**.

Objetivo 2

- O modelo a ser proposto visa o desenvolvimento de estratégias que vislumbrem a integração dos dados **na nuvem em níveis de instancia.**

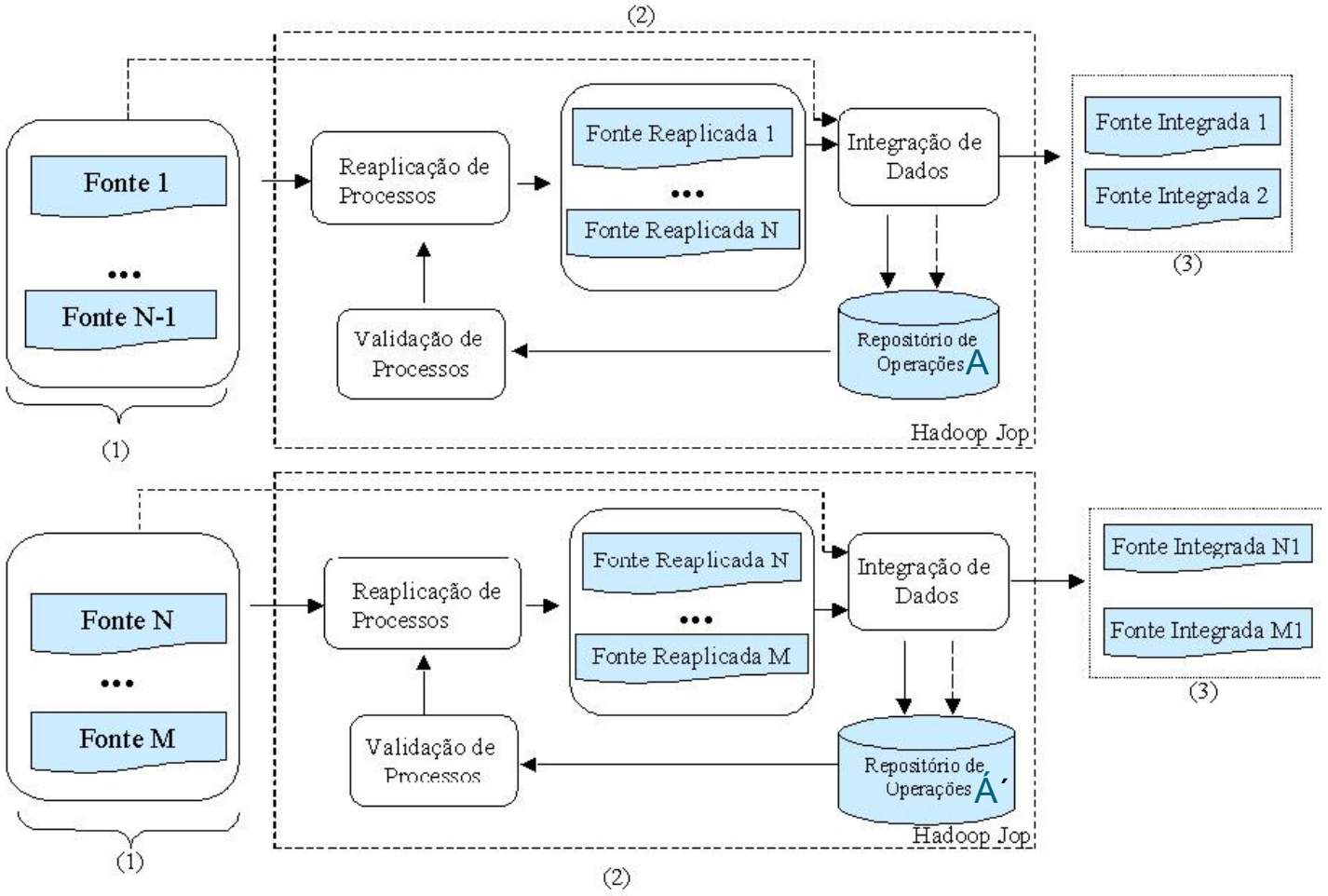
Hipótese 1:

“O modelo PrIntCloud **garante um bom desempenho e provê escalabilidade** na coleta de dados e no processamento de operações para a integração de dados em nível de instância considerando volumes de dados, mantendo **a qualidade**”

Hipótese 2:

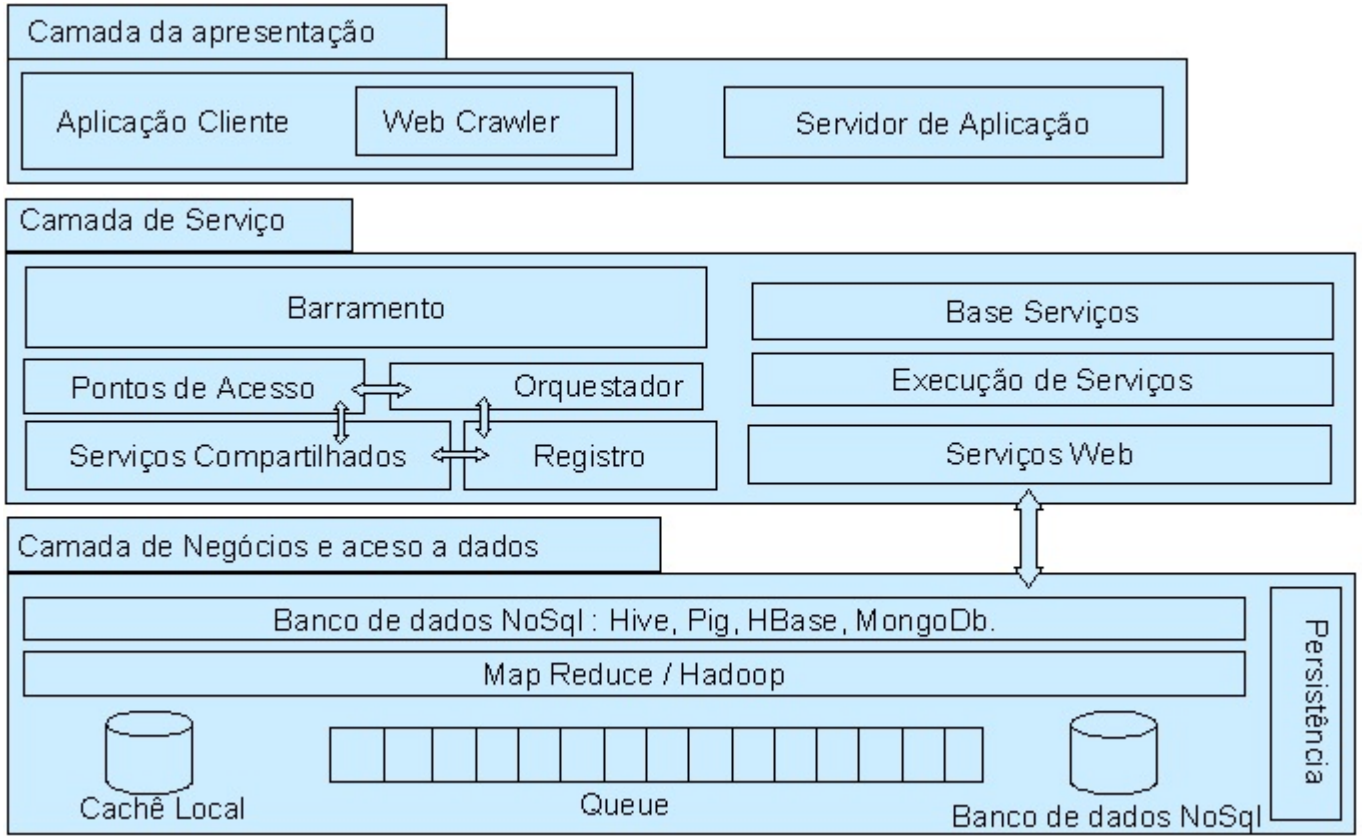
- “Uma abordagem PrIntCloud que combine técnicas de procedência de dados e integração de dados em nível de instância **reduz o tempo de processamento e o esforço humano** necessário para a integração de dados(quando se restabelece os resultados de processos de integração que já foram executados pelo menos uma vez)”.

Visão Geral do Modelo PrIntCloud



Repositório de operações A e Á são iguais.

Uma Arquitetura Escalável para Integração de Dados na Nuvem



3.-Desafios.

Desafios

I.-Propor formas distintas para estender o repositório do modelo PrInt de acordo com diferentes **paradigmas de armazenamento de dados em nuvem(NoSql)**.

esafios

II.-Propor uma **arquitetura escalável** para o processamento distribuído de *grandes volumes de dados* que combine técnicas que permitam integrar dados dos currículos Lattes com suporte à reaplicação de decisões anteriores;

Desafios

III.-Definir como os dados integrados em processos de integração anteriores são armazenados para serem acessados por processos de **integração subsequentes concorrentes na nuvem**. Assim, as transações podem ser armazenadas em um repositório centralizado;

Desafios

IV.- Este projeto de doutorado PrIntCloud tem como objeto propor o modelo distribuído na nuvem com acesso multiusuário na web.

Referências

Tomazela, B. (2010). MPPI: um modelo
de procedência para subsidiar processos de integração.
Dissertação de mestrado, Universidade de São Paulo, São
Carlos, SP.

Alfredo Goldman, Fabio Kon, Francisco Pereira Junior, Ivanilton Polato e Rosangela de Fátima Pereira disponivel BDBCOMP *http://www.lbd.dcc.ufmg.br/colecoes/jai/2012/003.pdf*, Universidade de São Paulo - IME

Livros

- Hadoop-The Definitive Guide(Tom White-2 Ed.)
- Hadoop in Action by Chuck Lam-1 Ed.
- web: http://wiki.apache.org/hadoop