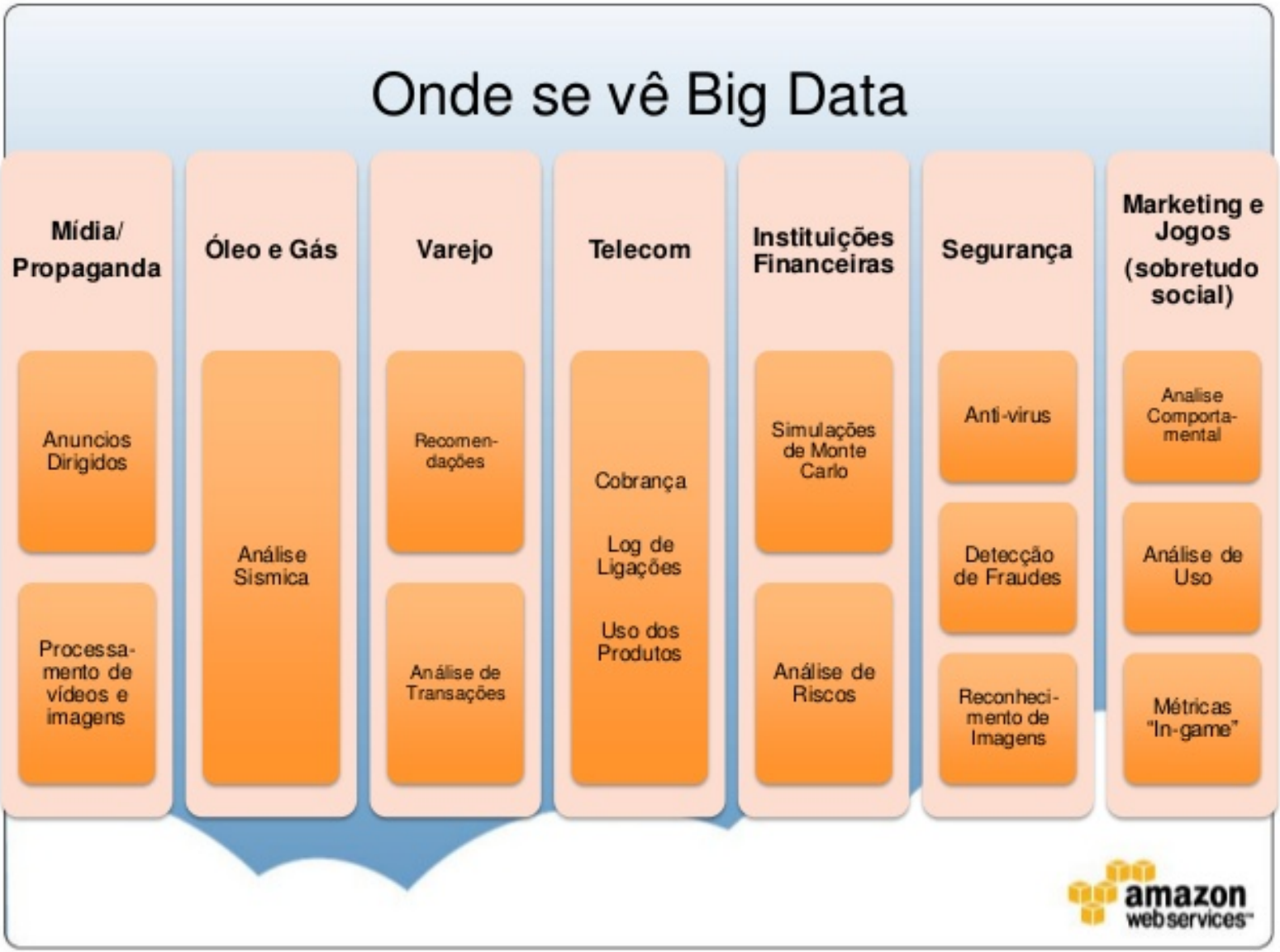
Hadoop

por
Guilherme Bayer Schneider

Dezembro - 2018 - Pelotas/RS


BIG
DATA
INFORMATION
STORAGE
LARGE
SETS
MILLION
SCIENCE
SOFTWARE
SAN
NAS
GOVERNMENT
PROCESS
VOLUME
TECHNOLOGIES
MANAGEMENT
PETABYTES
ANALYSIS
RESEARCH
INTERNET
SEARCH
LEARNING
EVERY
WORLD
BASED
DEVELOPMENT
TECHNOLOGY
BILLION
NEW
FUTURE
PARALLEL
INITIATIVE
SIZE
EXABYTES
FRAMEWORK
BUSINESS
DATABASES
PROCESSING
ANALYTICS
SIMULATIONS
APPLICATIONS
AMOUNT
REAL
STRUCTURE
WORK
INFRASTRUCTURE
SOCIAL
YEARS
RESULTS
DRIVE
REPLICATION
HANDLE
SOURCES
CHALLENGES
LIMITS
COMPLEX
PAST
ANNOUNCED
HAND
CAPACITY
ONE
OUT
INCLUDE
NETWORKS
HYPOTHESIS
END
MOST
NEARLY
MAPREDUCE
CRITIQUES
STATE
CRITIQUE
PEOPLE
TRAFFIC
USE
CURRENTLY
APPROACH
MANY
FACTOR
ONLINE
DAY
WORLDWIDE
SYSTEMS
MANAGE
MAKING
INTEGRATION
MASSIVE
DEFINITION
BETWEEN
USING
STATISTICS
EFFECTIVE
VISUALIZATION
DETERMINE
WELL
SET
DEPARTMENT
QUERIES
HIGH
USERS
HIGHER
ALGORITHMS
COMPANIES
RATE
TERMS
PRIVATE
YEAR
CAPABILITIES
SUCCESSFUL
NEEDED
USES
COMPARED
SHARED
LESS
HUGE
VARIOUS
TIMES
ARCHITECTURE
PRIVACY
INTERNATIONAL
IMPLEMENTATION
BECOMES
SECOND
NATIONAL
FLOW
INTELLIGENCE
APPROACHES
PROCESSES
PARADIGM
FUNDING
INSIGHT
TOOLS
CENTER
RECORDS
COST
PER
DISTRIBUTED
RELEVANT
USED
PROJECT
ANNUAL
NOW
ECONOMIC
MARKET
VARIETY
SENSOR
TIME
WOULD
TERABYTES
RELATED

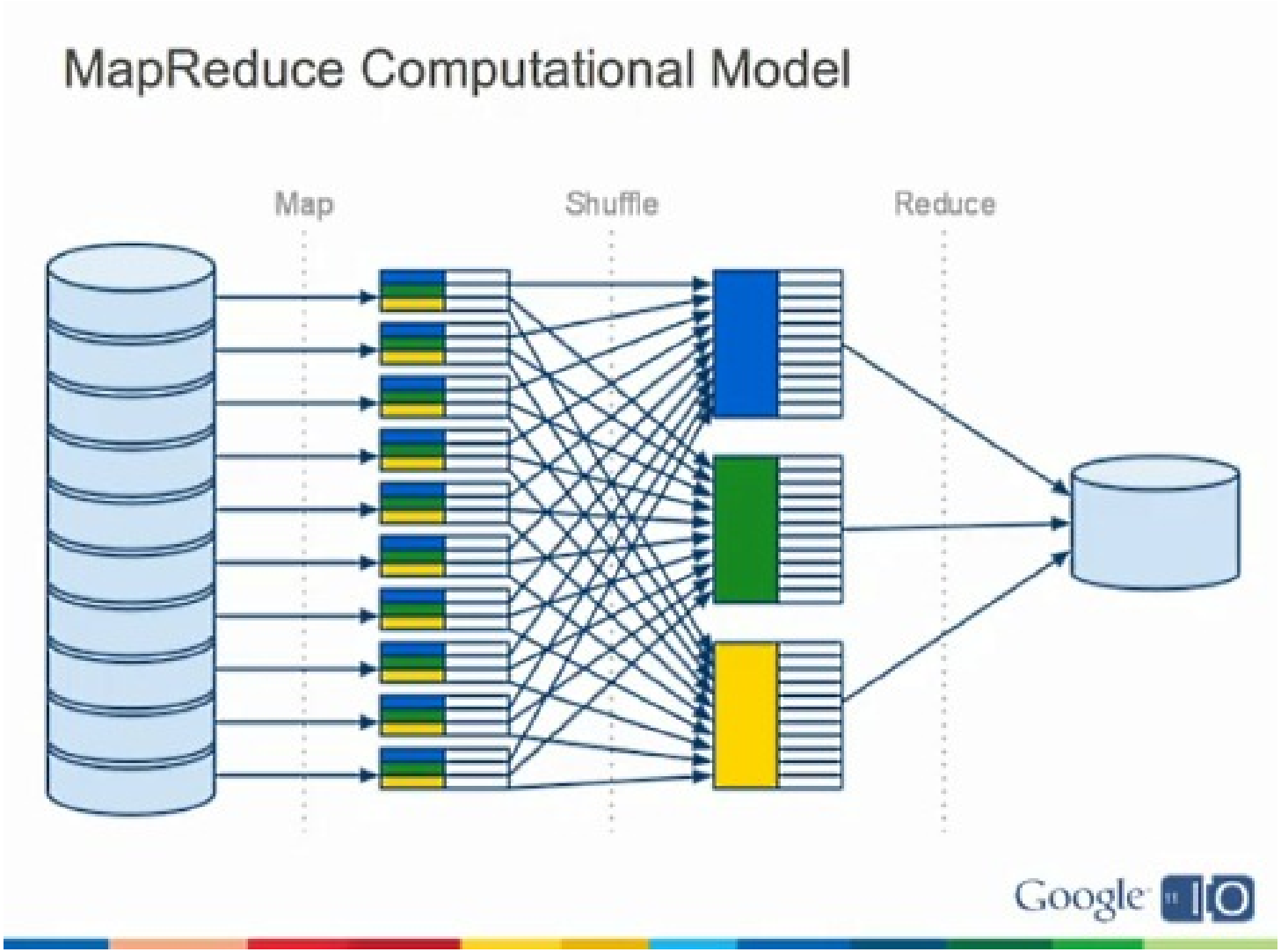

Onde se vê Big Data
Mídia/ Propaganda
Anuncios Dirigidos
Processa-mento de vídeos e imagens
Óleo e Gás
Análise Sísmica
Varejo
Recomen-dações
Análise de Transações
Telecom
Cobrança
Log de Ligações
Uso dos Produtos
Instituições Financeiras
Simulações de Monte Carlo
Análise de Riscos
Segurança
Anti-virus
Detecção de Fraudes
Reconheci-mento de Imagens
Marketing e Jogos (sobretudo social)
Analise Comporta-mental
Análise de Uso
Métricas "In-game"
amazon web services

Big Data no Marketing

- Série House of Cards
  - A partir da análise de navegação e reviews, Netflix percebeu que poderia criar uma série de sucesso;
  - Viu que Kevin Spacey tinha grande aceitação a partir da análise de dados;
  - Entendeu que thrillers políticos tinham grande apelo com o seu público;
- Recomendação de filmes
  - Utiliza machine learning com técnicas de filtro colaborativo e *k-nearest neighbor*.
  - Recomenda filmes com precisão com um catálogo gigantesco. No Brasil o catálogo é menor, mas ainda assim a precisão é muito boa.

MapReduce

- Aplicação que facilita a programação de tarefas a grandes quantidades de dados(multi-terabyte data-sets).
- Realiza computação sobre os dados (pouca movimentação de dados).
- Utiliza os blocos armazenados no HDFS, logo não necessita de divisão dos dados.


MapReduce Computational Model
Map
Shuffle
Reduce
Google I/O

O Hadoop

- Aplicação para processamento e armazenamento de dados em larga escala;
- Implementado em Java;
- É integrado com MapReduce, permitindo que os dados sejam computados localmente quando possível;
- Tecnologia recente, porém já muito usada.

Vantagens(1)

- Menos suscetível a erros, pois não é concentrado em uma máquina;
- Código Aberto;
- Aluguel de serviços disponíveis na nuvem;
- Uso de hardware convencional;

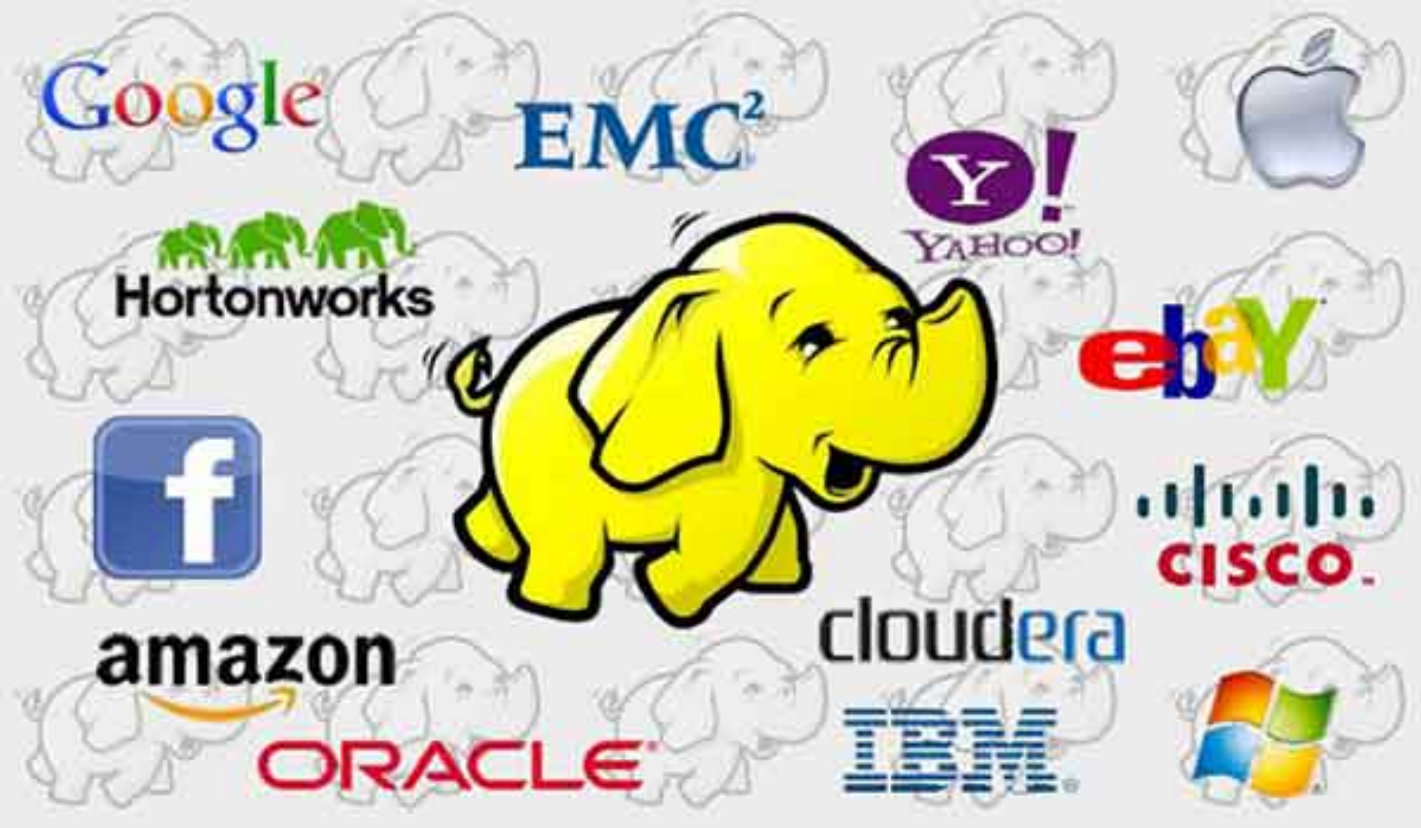
Vantagens(2)

- A falha individual de máquinas não afeta a disponibilidade dos dados;
- Há a possibilidade de manipulação de dados localmente;
- Divisão de subtarefas;
- Balanceamento de carga;

Desvantagens

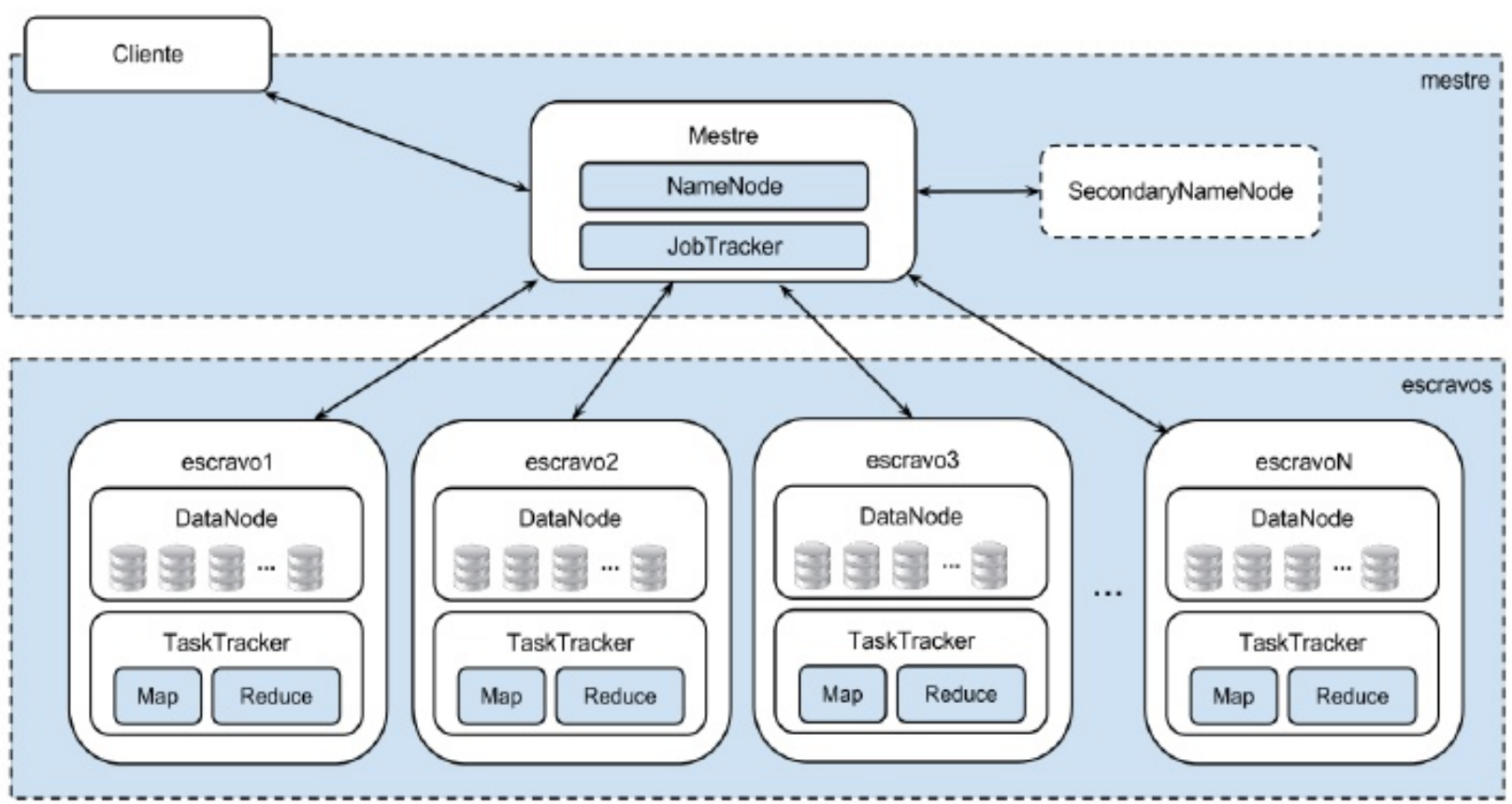
- Nó mestre é ponto único de falha;
- Como o nó mestre é único pode impedir o escalonamento;
- Dificuldade no processamento de arquivos pequenos;
- Baixo desempenho em aplicações que necessitem de muito processamento em pequenos arquivos;
- Não é designado para aplicações que necessitem de baixa latência(milisegundos);

Algumas empresas que atualmente utilizam o Hadoop



http://cdn.edureka.co/blog/wp-content/uploads/2013/04/2-o.jpg

Distribuição dos componentes do Hadoop

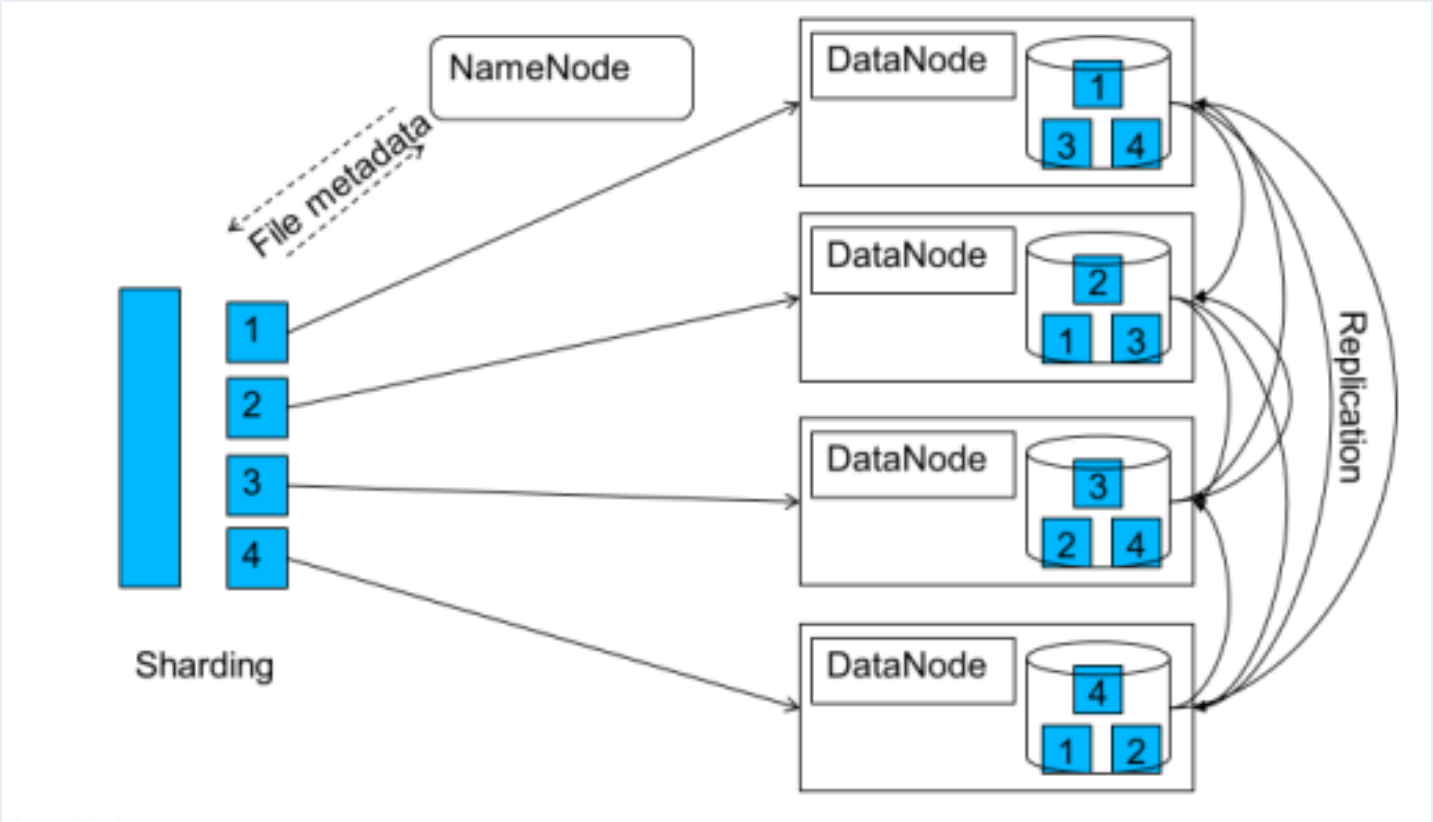


http://www.ime.usp.br/ ipolato/JAI2012-Hadoop.pdf

Hadoop Distributed File System

- Projetado para grandes quantidades de dados(terabytes ou petabytes);
- A falha individual de máquinas não afeta a disponibilidade dos dados;
- Garantia de redundância;
- Alto desempenho na manipulação de grandes arquivos;
- É integrado com MapReduce, permitindo que os dados sejam computados localmente quando possível;
- É projetado para que o arquivo seja escrito uma única vez e lido milhares de vezes.

Diagrama de replicação executado pelo Hdfs



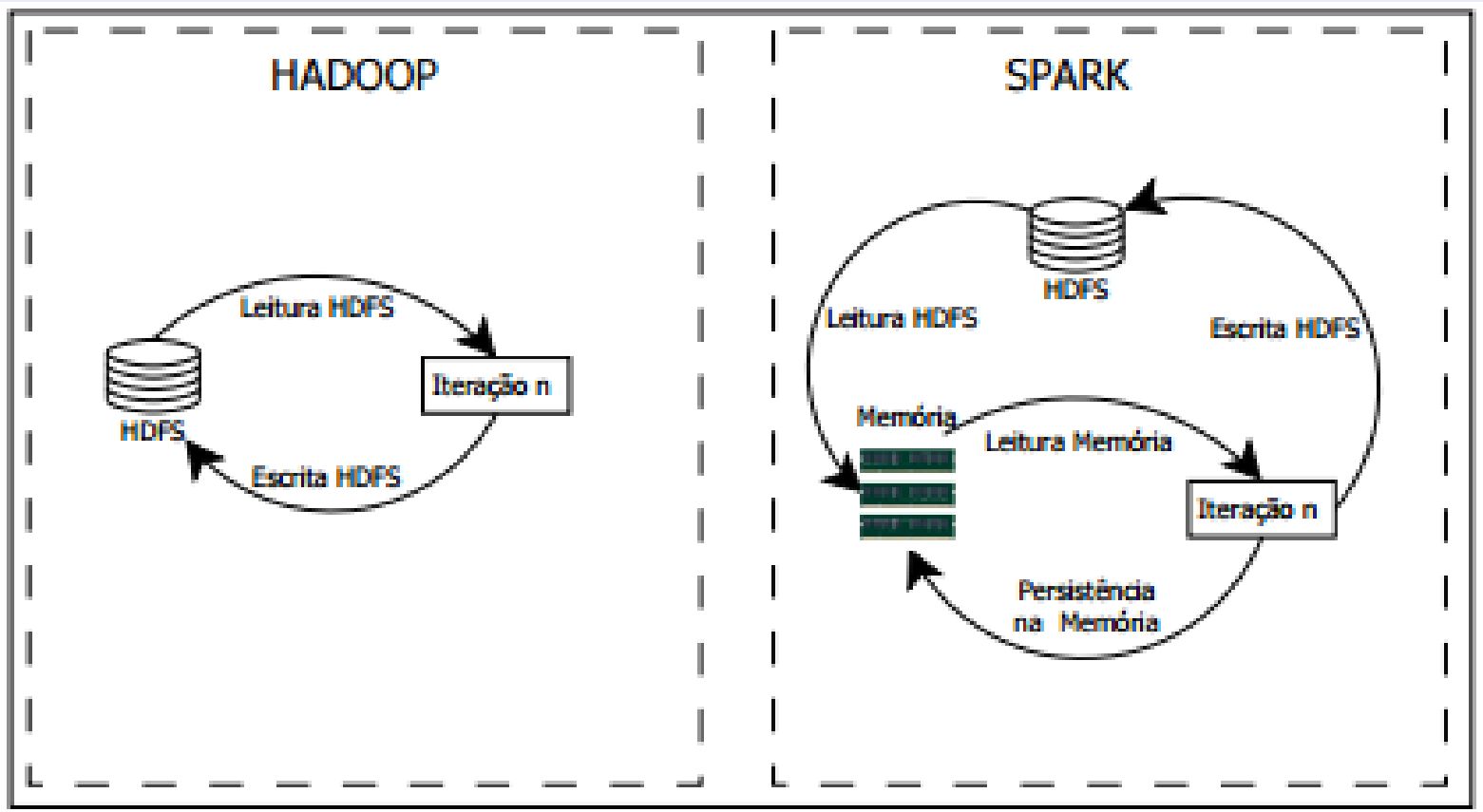
https://cvw.cac.cornell.edu/mapreduce/images/hdfs.png

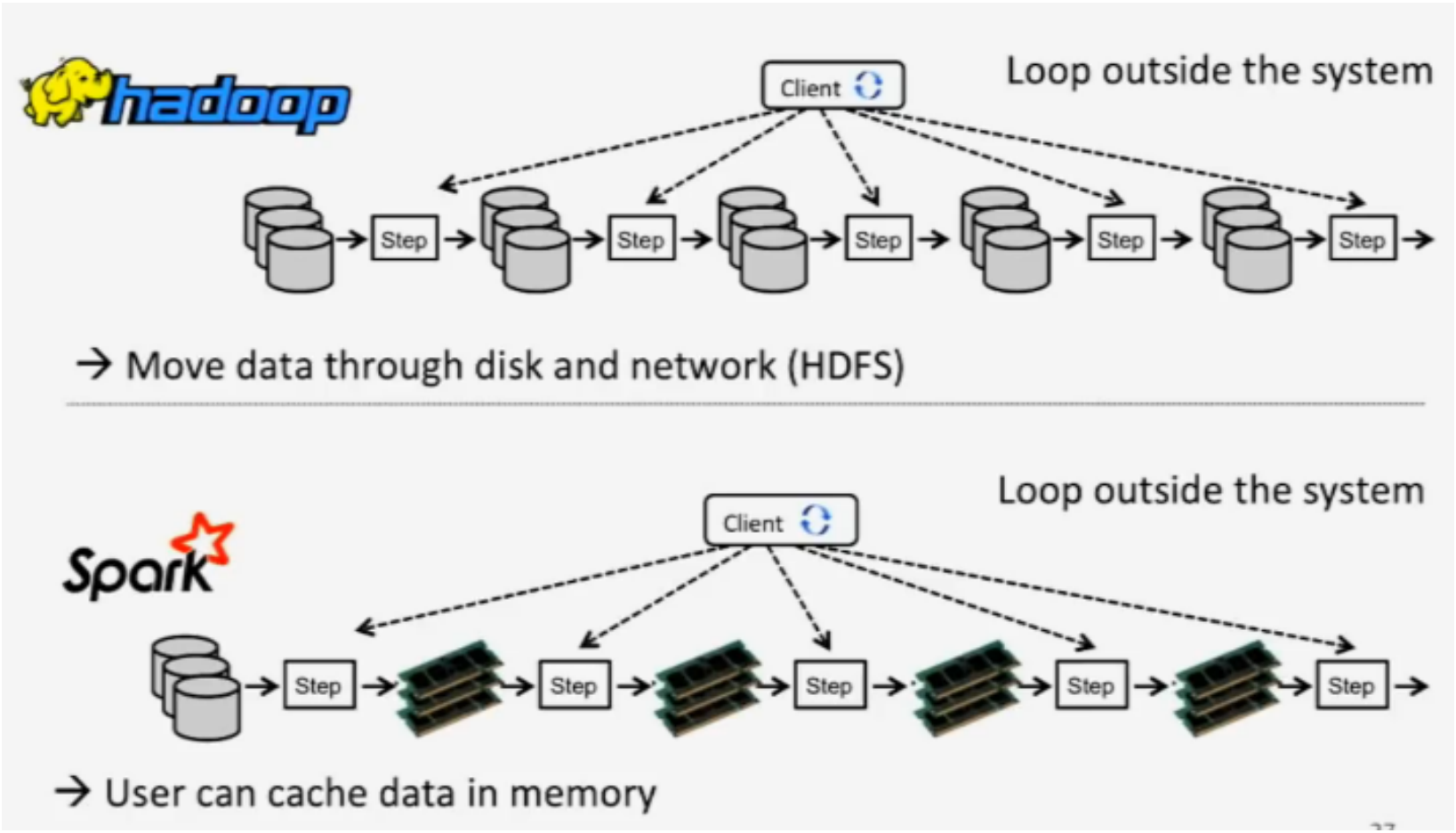
Projetos Relacionados ao Hadoop

Apache Spark

- Possibilita o uso de memória através da tecnologia Resilient Distributed Datasets(RDDs).
- É otimizado para baixa latência.
- Consegue executar tarefas abaixo de 100 milisegundos enquanto que o Hadoop demoraria de 5 a 10 segundos.

Apache Spark





http://www.nextplatform.com/wp-content/uploads/2015/02/flinkIterative.png

- **Ambari™**:Ferramenta para administração de Hadoop Clusters.
- **Cassandra™**:Banco de dados multi-master sem ponto único de falha.
- **Chukwa™**:Sistema de análise da dados coletados.
- **HBase™**:Sistema distribuído que suporta armazenamento de grandes tabelas.
- **Hive™**:É uma infraestrutura de Data warehouse e possibilita a manipulação dos dados por meio de uma linguagem muito semelhante ao Sql que é o HiveQl.
- **Pig™**:Oferece uma linguagem de processamento de dados de alto nível.
- **Tez™**:Ferramenta para auxílio em operações mapreduce.

Como grandes empresas o utilizam?

- **Adobe (www.adobe.com)**:Onde usa: no armazenamento e processamento de dados internos e de redes sociais. Parque:aproximadamente 80 nós de processamento.
- **e-Bay (www.ebay.com)**:Onde usa: na otimização de buscas. Parque: aproximadamente 532 nós de processamento.
- **Facebook (www.facebook.com)**:Atualmente conta com mais de 845 milhões de usuários ativos.Onde usa: análise de log. Parque: aproximadamente 1.400 nós de processamento
- **LinkedIn (www.linkedin.com)**:Onde usa: análise e busca de similaridade entre perfis de usuários. Parque: aproximadamente 1.900 nós de processamento.
- **Twitter (www.twitter.com)**:Onde usa: no armazenamento de mensagens e no processamento de informações. Parque: não divulgado.
- **Yahoo! (www.yahoo.com)**:Onde usa: no processamento de buscas, recomendações de publicidades, testes de escalabilidade. Parque: aproximadamente 40.000 nós de processamento.

Conclusão

- Possibilita o armazenamento de dados na escala petabyte;
- Processamento local de dados economizando uma quantidade significativa de tráfego;
- Garantia de redundância;
- Não designada para aplicações que necessitem de baixa latência(milisegundos);