



Engenharia de Dados com Hadoop e Spark



Bem-vindo(a)





Planejando e Configurando um Cluster Hadoop

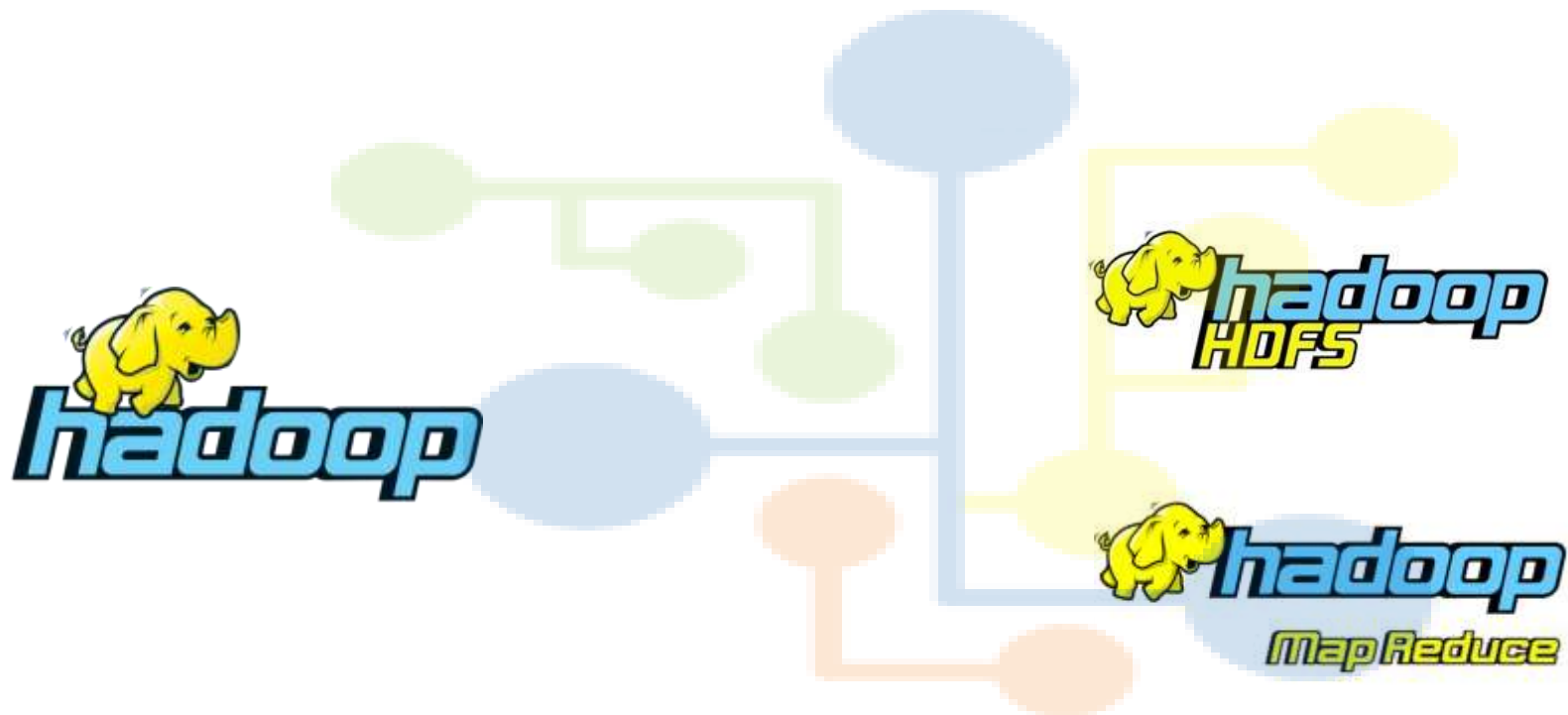
A faint, stylized diagram of a Hadoop cluster topology is visible in the background. It shows a central node connected to several other nodes, with lines representing network connections. The nodes are colored in shades of blue, green, yellow, and orange, matching the Data Science Academy logo.

Cluster Hadoop



Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d



Cluster Hadoop



Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d



- Arquitetura de um Cluster Hadoop
- Topologia de Rede para o Cluster Hadoop
- Workflow
- Planejamento do Cluster
- Hardware Configuração de Rede do Cluster Hadoop
- Arquivos de Configuração
- Parâmetros de Configuração
- Como funciona o HDFS
- HDFS Writes
- HDFS Reads
- Importando Dados do MySQL para o HDFS

Cluster Hadoop



Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

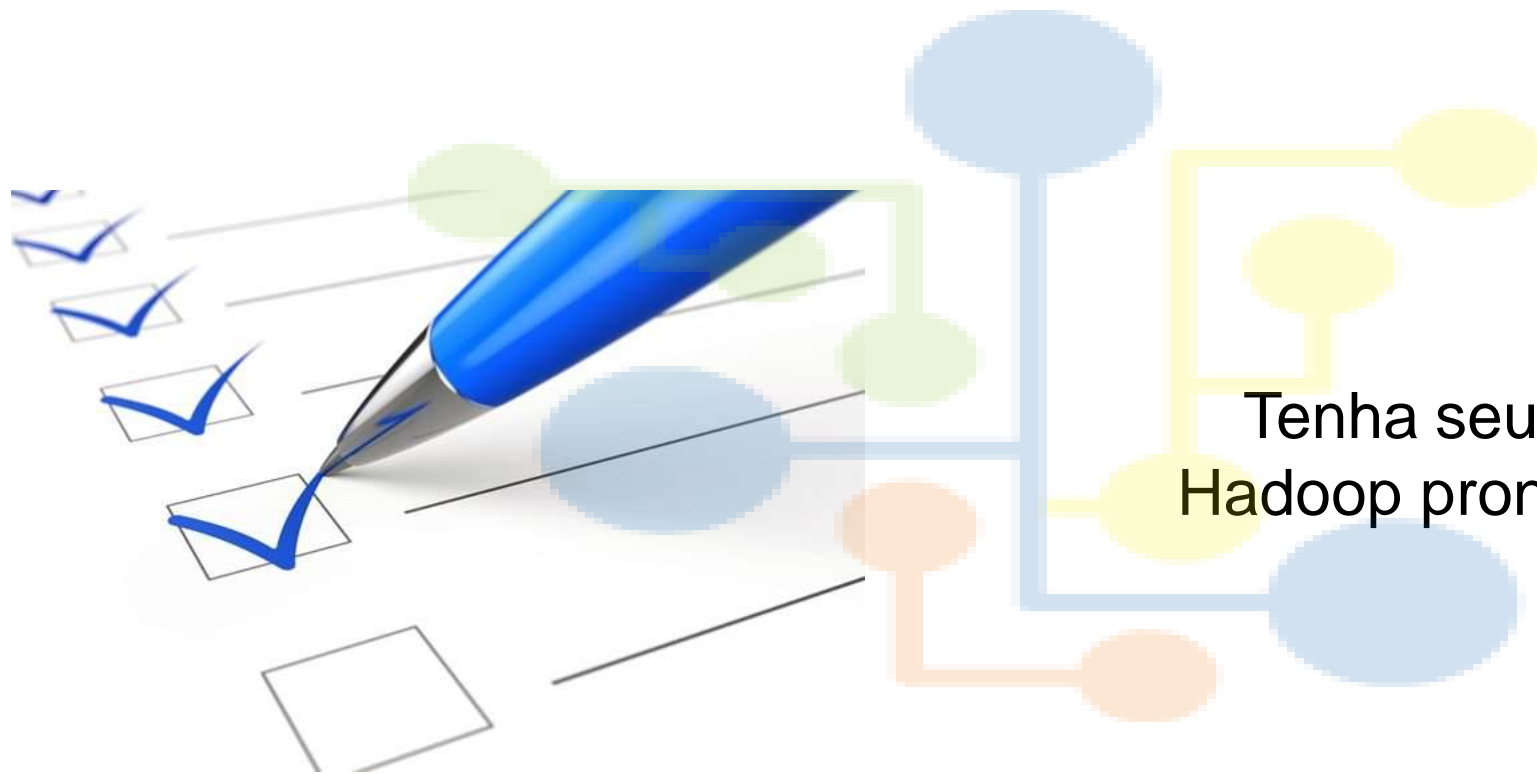


Muito do que vamos estudar neste capítulo é responsabilidade do Engenheiro de Dados.

É este profissional que deve entregar ao Cientista de Dados a infraestrutura necessária para análise de Big Data.



Cluster Hadoop



Tenha seu ambiente
Hadoop pronto para uso.



O que é um Cluster?





O que é um Cluster?



Cluster de Computadores



O que é um Cluster?

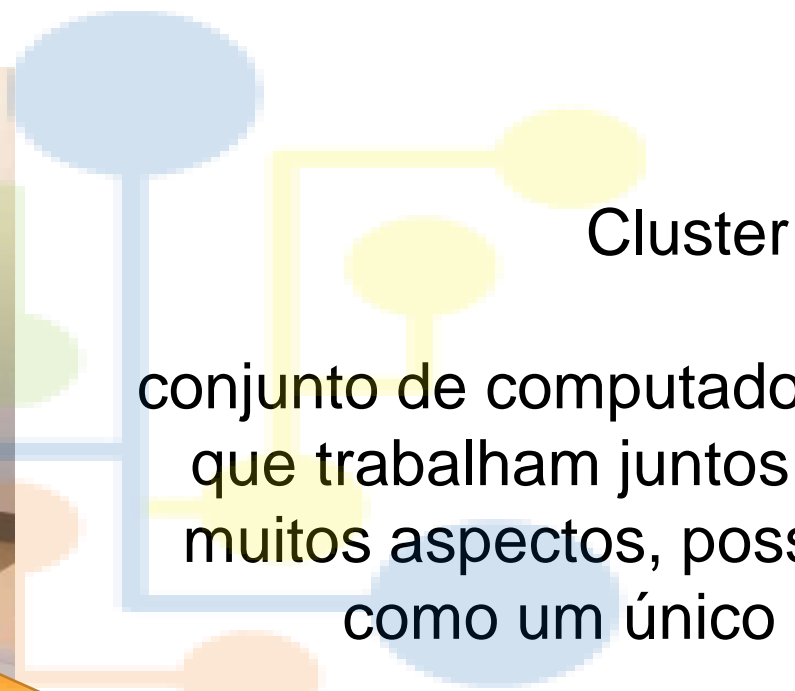
A network diagram showing several nodes (circles) connected by lines. The nodes are colored in blue, green, yellow, and orange. The word 'Cluster' is written in the center of the diagram.

Cluster

Um cluster é um conjunto de computadores conectados que trabalham juntos para que, em muitos aspectos, possam ser vistos como um único sistema. Os clusters de computadores têm cada nó configurado para executar a mesma tarefa, controlada e programada por software.



O que é um Cluster?



Cluster

conjunto de computadores conectados que trabalham juntos para que, em muitos aspectos, possam ser vistos como um único sistema.

Node



O que é um Cluster?



Cluster

conjunto de computadores conectados que trabalham juntos para que, em muitos aspectos, possam ser vistos como um único sistema.



O que é um Cluster?





O que é um Cluster?

- Cluster de Alto Desempenho

Existem diversos tipos de Cluster



1 gigaflop corresponde a 1 bilhão de instruções por segundo



O que é um Cluster?

- Cluster de Alto Desempenho
- Cluster de Alta Disponibilidade
- Cluster para Balanceamento de Carga
- Cluster Combo

Existem diversos tipos de Cluster



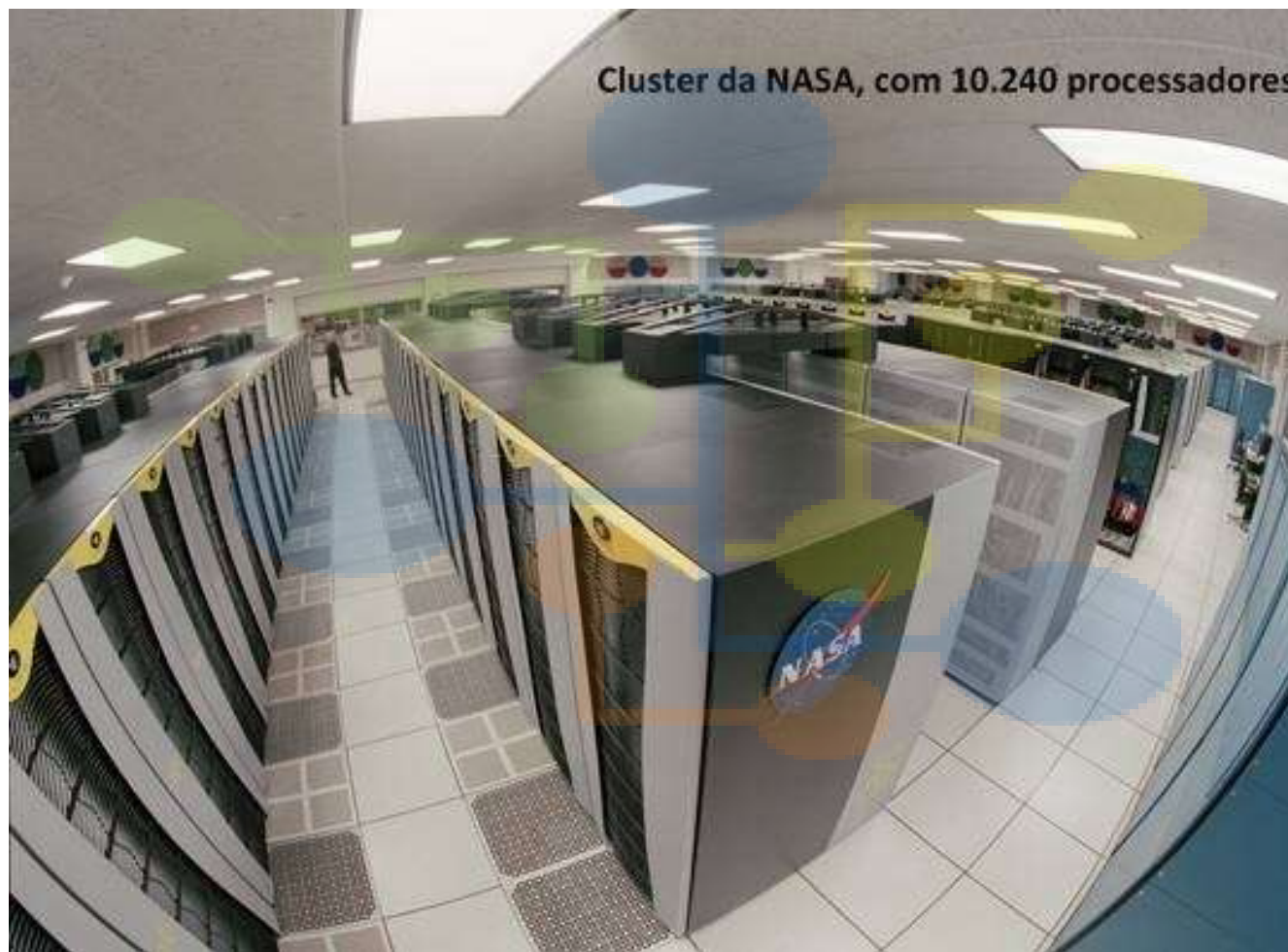


O que é um Cluster?

Para uma aplicação de Big Data podemos configurar um cluster de alto desempenho e ao mesmo tempo alta disponibilidade, se for necessário processamento e análise de dados em tempo real, para um sistema de recomendação, por exemplo.



O que é um Cluster?





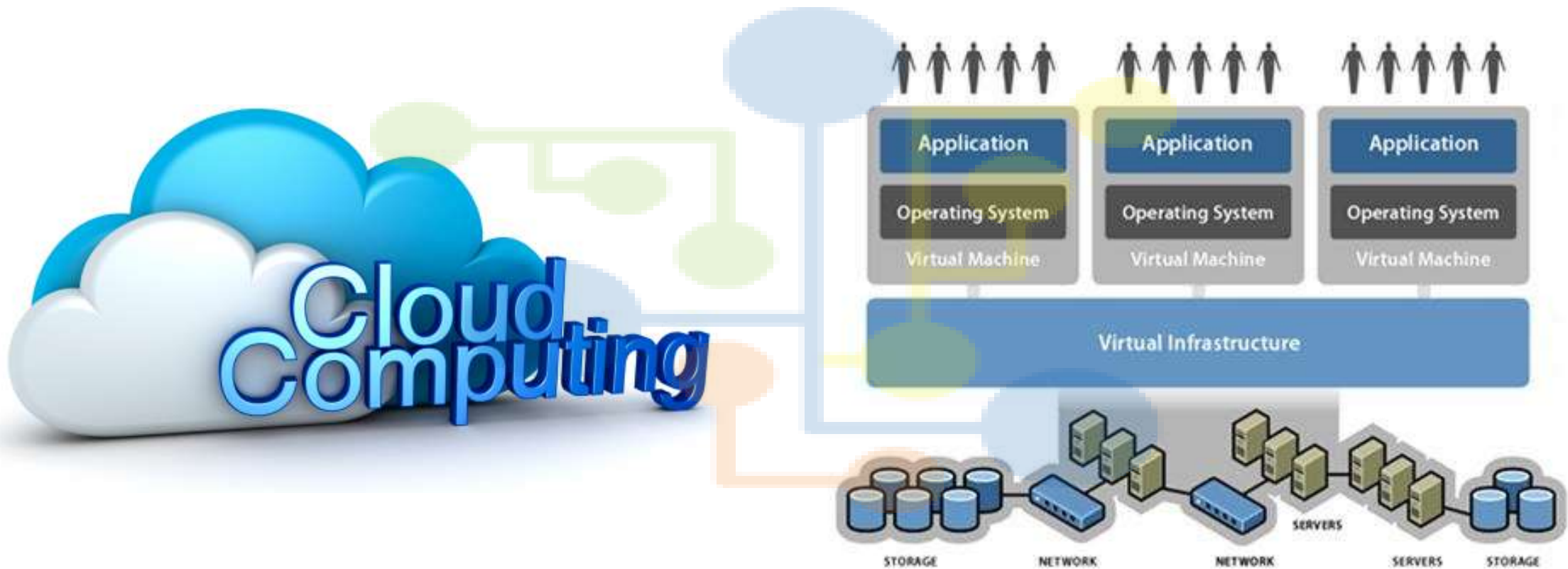
O que é um Cluster?



As tecnologias de Clustering possibilitam a solução de diversos problemas que envolvem grande volume de processamento.



O que é um Cluster?





Arquitetura do Cluster Hadoop





Arquitetura do Cluster Hadoop

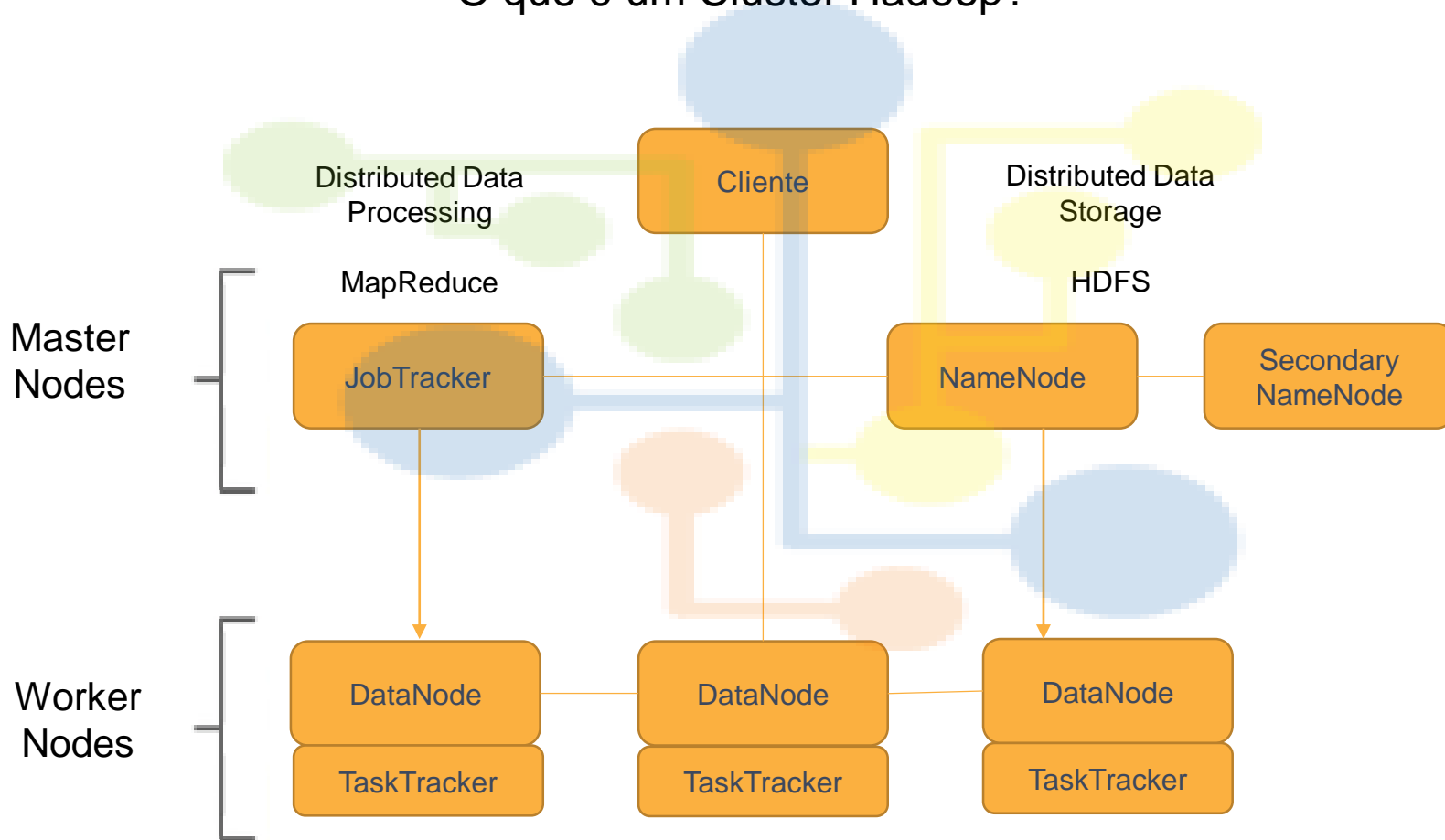
O que é um Cluster Hadoop?

Um Cluster Hadoop é um conjunto de máquinas com Hadoop instalado que é criado para armazenar e analisar grandes quantidades de dados, sejam eles estruturados ou não **estruturados**. Em um Cluster Hadoop, os dados são armazenados e processados ao longo de diversos computadores e tudo isso é feito de forma paralela.



Arquitetura do Cluster Hadoop

O que é um Cluster Hadoop?





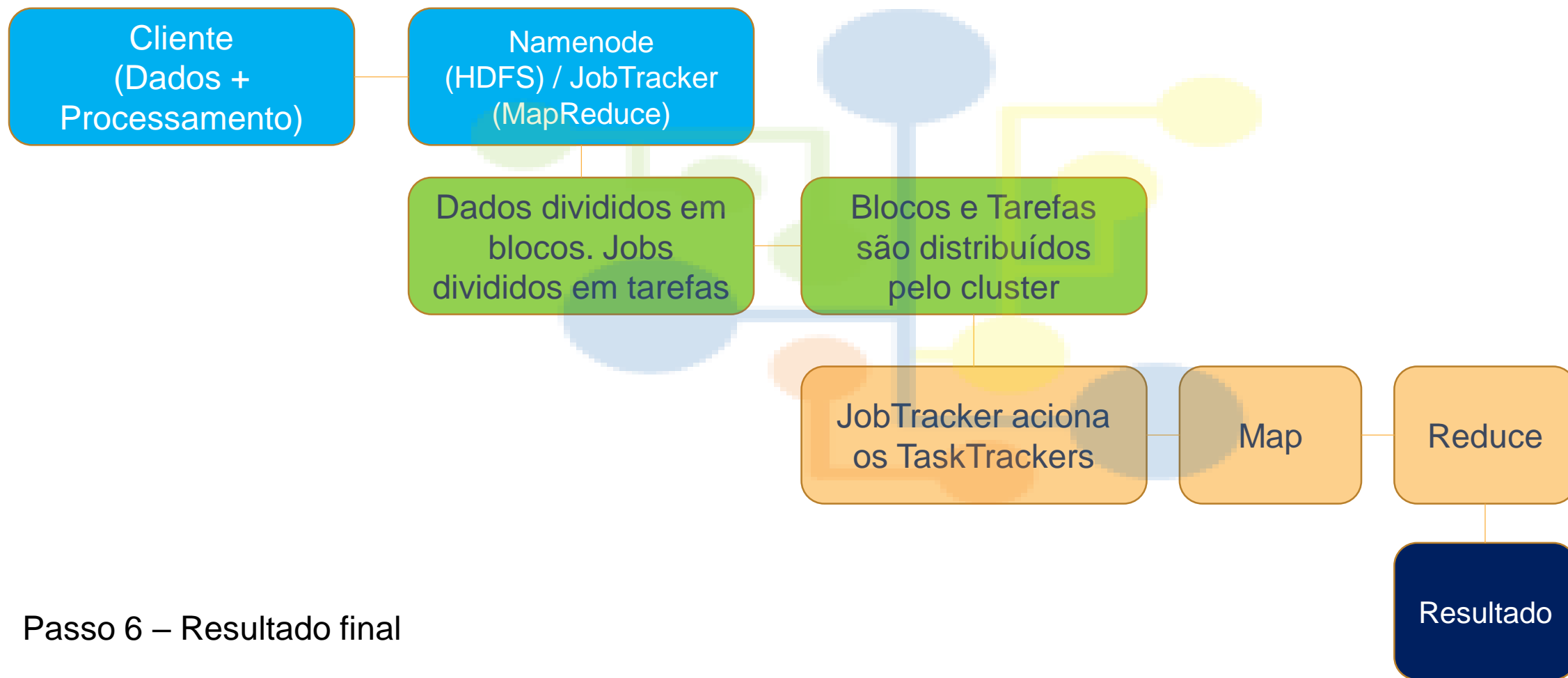
Arquitetura do Cluster Hadoop



Funcionamento do Cluster



Arquitetura do Cluster Hadoop

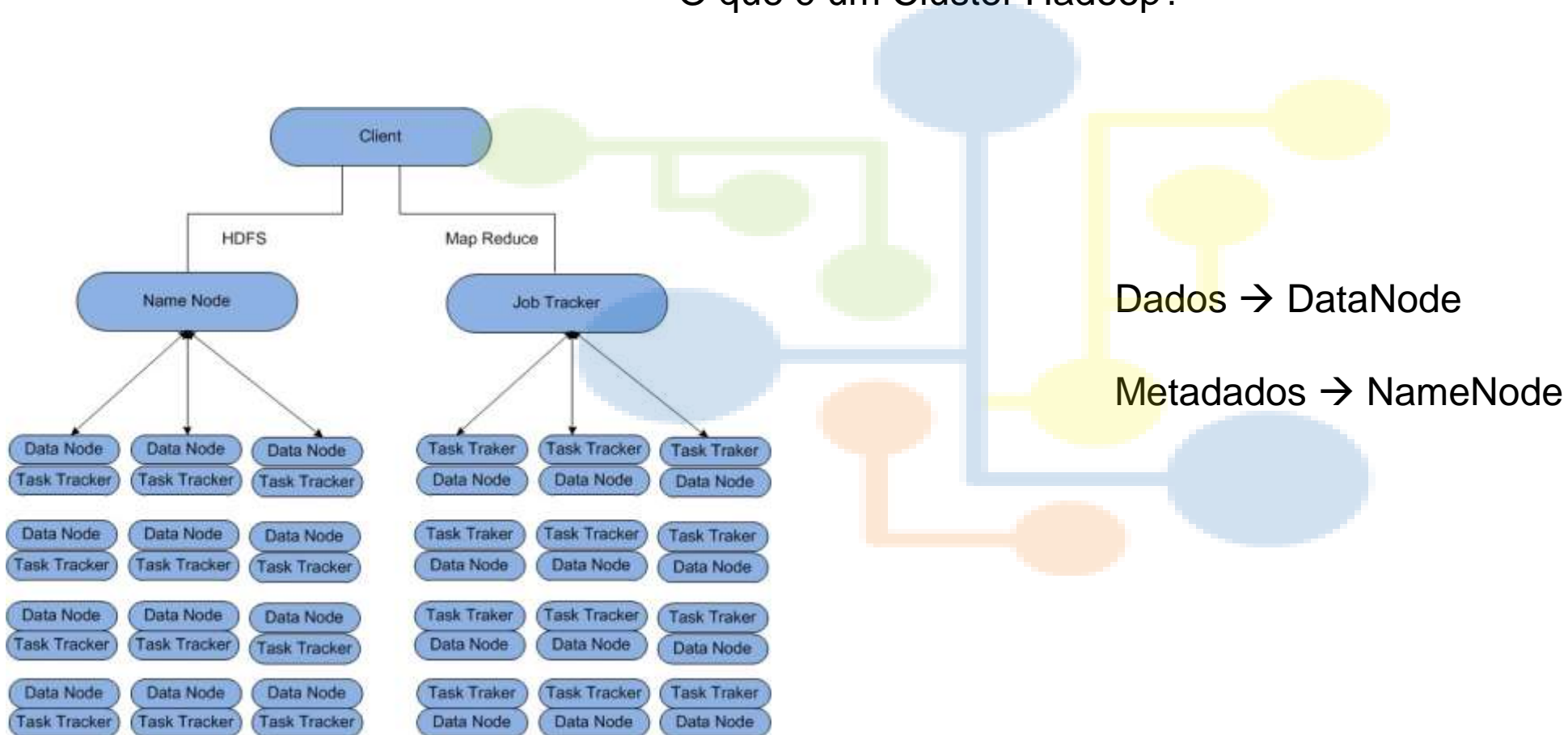


Passo 6 – Resultado final



Arquitetura do Cluster Hadoop

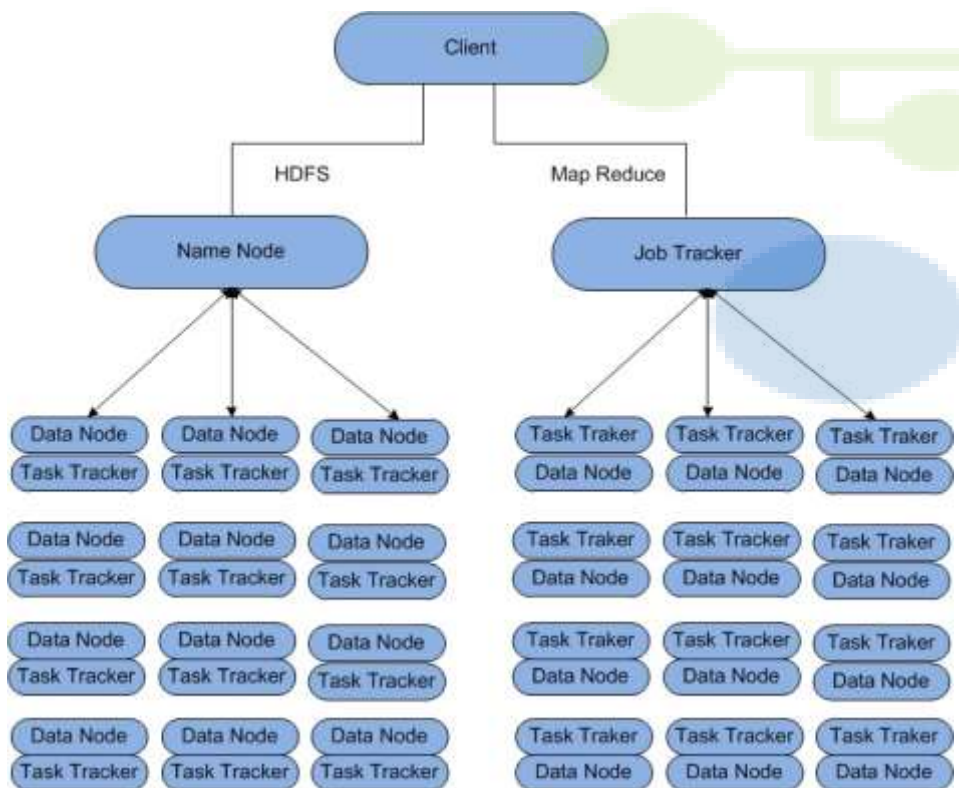
O que é um Cluster Hadoop?





Arquitetura do Cluster Hadoop

O que é um Cluster Hadoop?



DataNode → Armazena/Recupera Dados

TaskTracker → Executa Jobs de MapReduce



Topologia de Rede do Cluster Hadoop



Cluster Hadoop

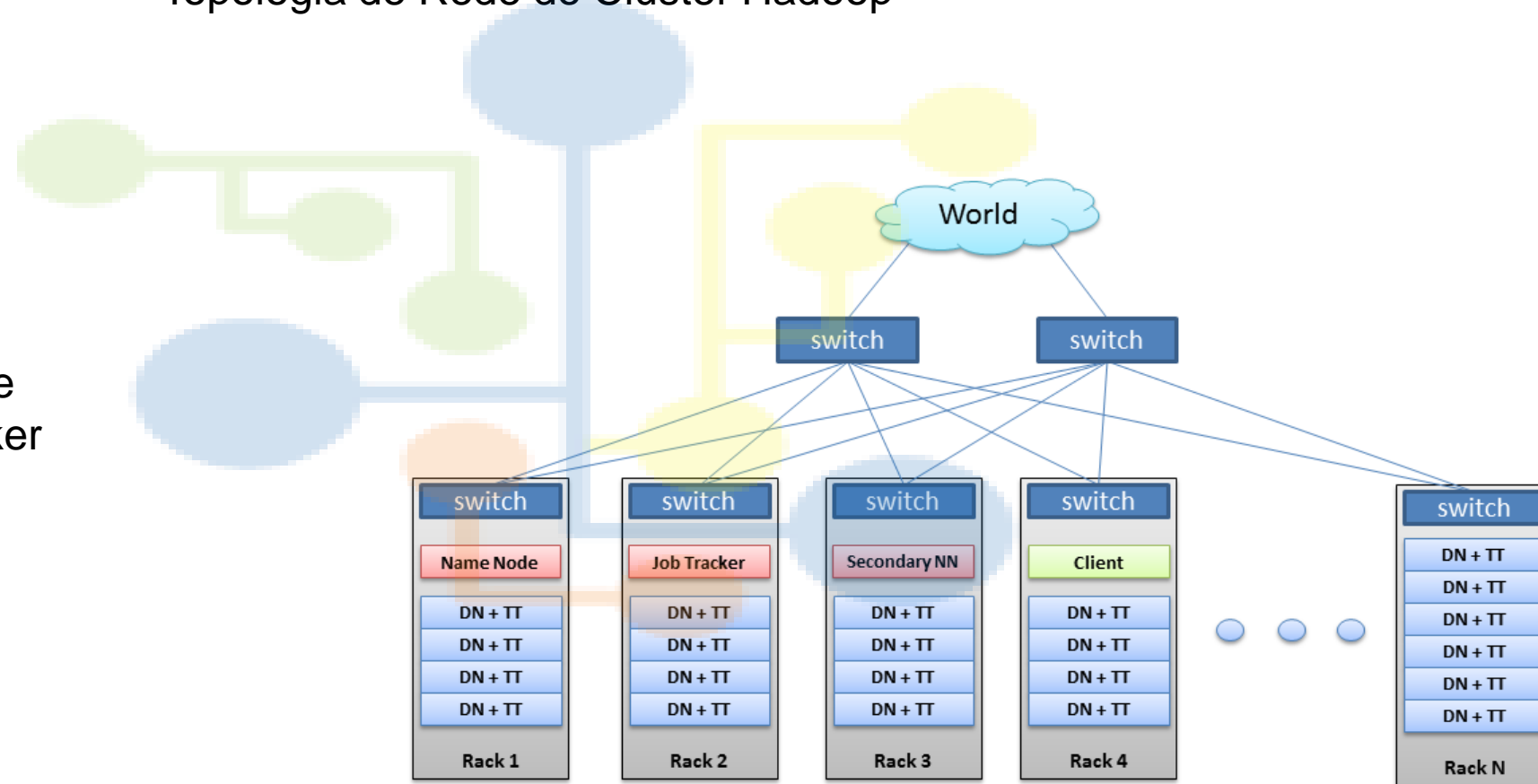


Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

Topologia de Rede do Cluster Hadoop

DN = DataNode
TT = TaskTracker



Cluster Hadoop



Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

Topologia de Rede do Cluster Hadoop



Rack



Switch

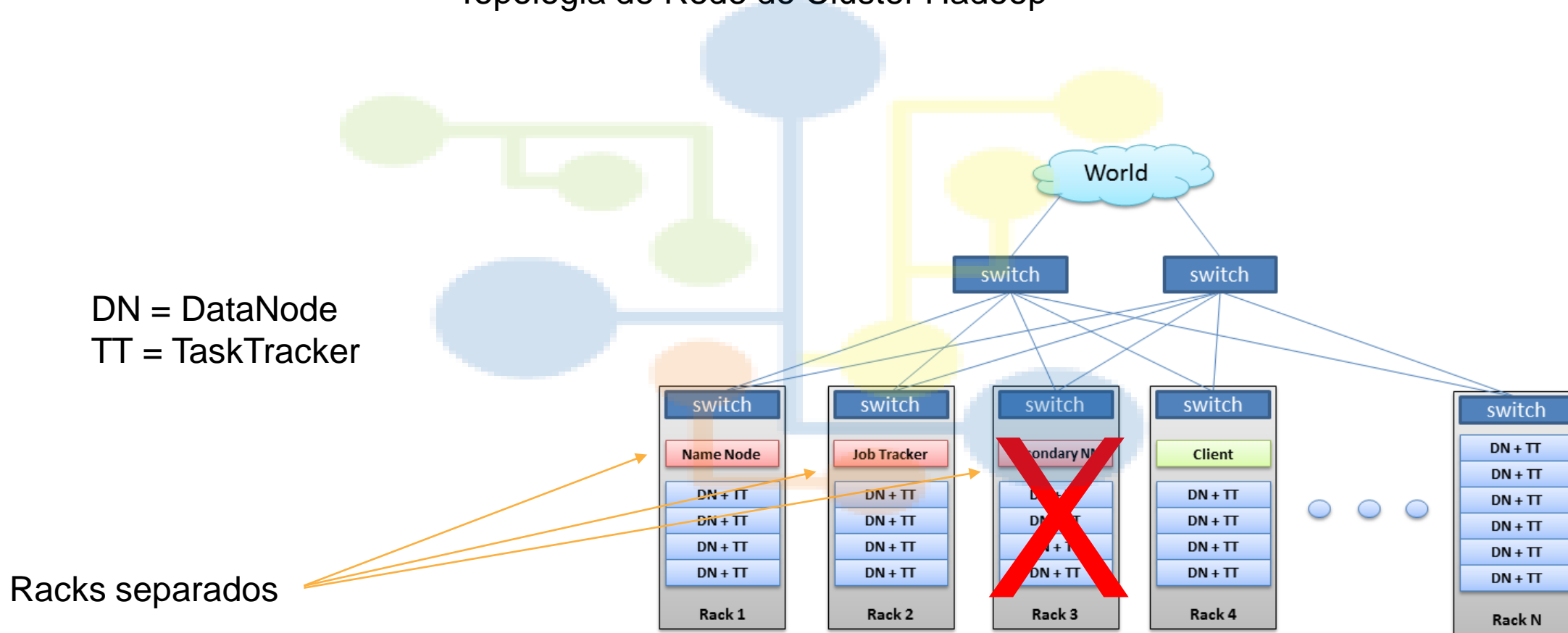
Cluster Hadoop



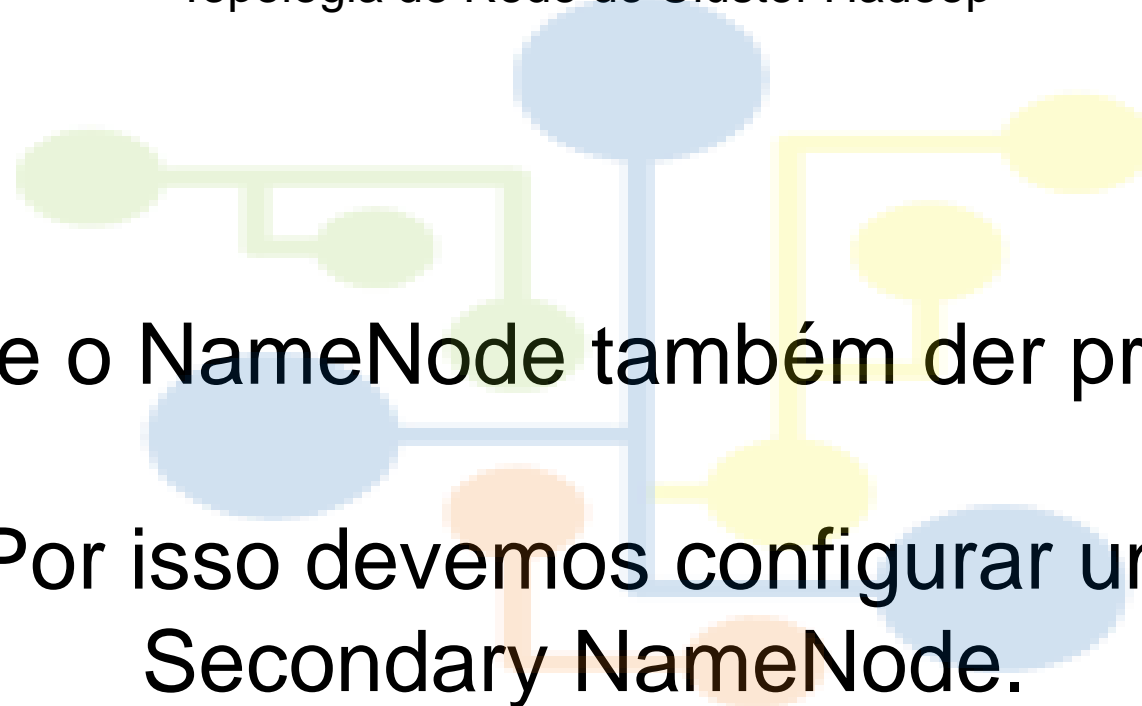
Data Science
Academy

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

Topologia de Rede do Cluster Hadoop



Topologia de Rede do Cluster Hadoop



Mas e se o NameNode também der problema?

Por isso devemos configurar um
Secondary NameNode.



Workflow do Cluster Hadoop





Workflow do Cluster Hadoop

Workflow de um Cluster Hadoop

Os dados são divididos em blocos e distribuídos pelo cluster Hadoop

MapReduce analisa os dados baseado nos pares de chave-valor

Os resultados são colocados em blocos através do cluster Hadoop

Os resultados podem ser lidos do cluster



Workflow do Cluster Hadoop

Workflow de Gravação de Dados no HDFS



O objetivo do Cluster Hadoop, é o rápido processamento, em paralelo, de grandes quantidades de dados.

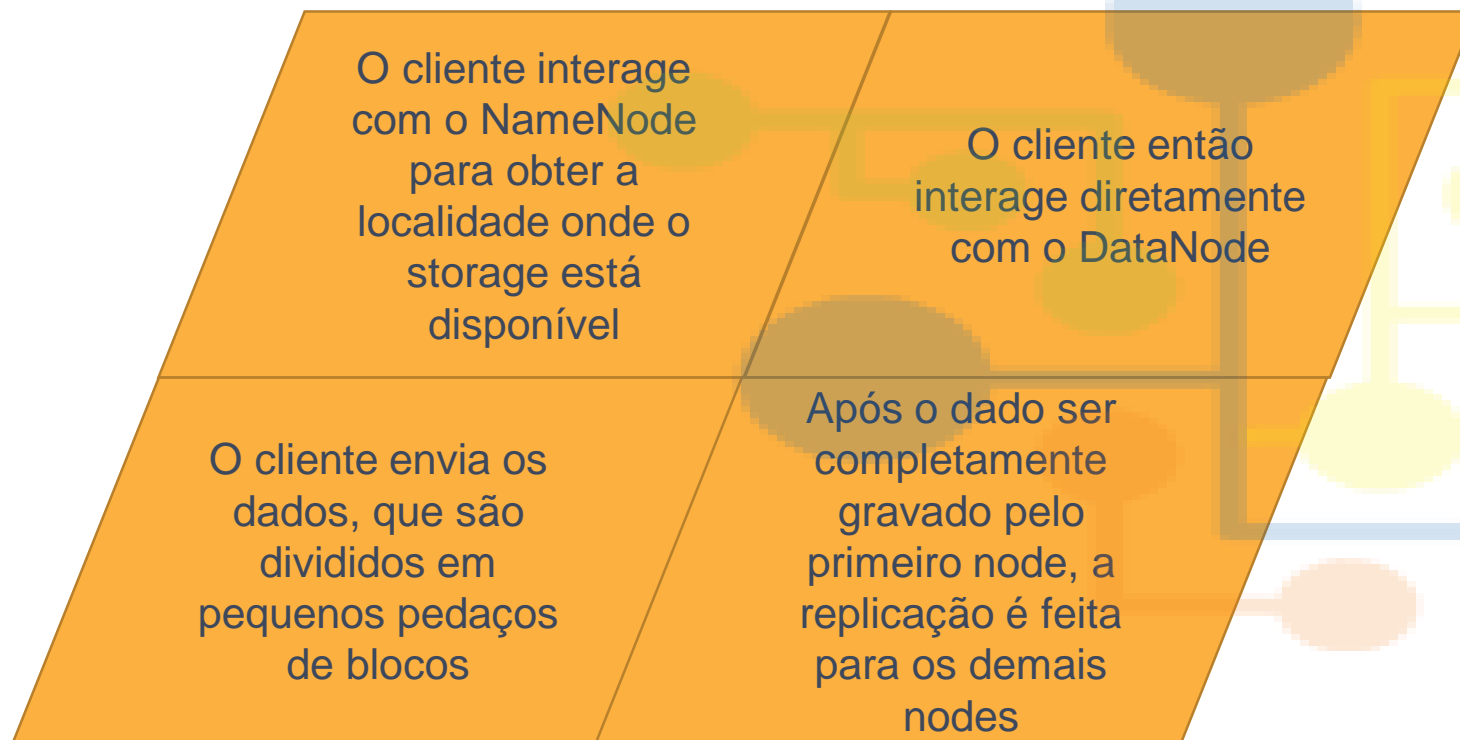
A configuração padrão do Hadoop, é ter 3 cópias de cada bloco de dados no cluster (o que pode ser modificado pelo parâmetro `dfs.replication` no arquivo de configuração `hdfs-site.xml`).

Vamos verificar, como é o processo de gravação de dados no HDFS.



Workflow do Cluster Hadoop

Workflow de Gravação de Dados no HDFS



O objetivo do Cluster Hadoop, é o rápido processamento, em paralelo, de grandes quantidades de dados.

A configuração padrão do Hadoop, é ter 3 cópias de cada bloco de dados no cluster (o que pode ser modificado pelo parâmetro `dfs.replication` no arquivo de configuração `hdfs-site.xml`).

Vamos verificar, como é o processo de gravação de dados no HDFS.



Workflow do Cluster Hadoop

Workflow de Gravação de Dados no HDFS


Após todos os DataNodes terminarem a gravação do dado, o relatório de blocos envia um sinal ao cliente, que então comunica o NameNode.

Os DataNodes também enviam o relatório de blocos ao NameNode.

O NameNode utiliza o relatório de blocos para atualizar os Metadados.



Workflow do Cluster Hadoop

A faint, stylized diagram of a Hadoop cluster is visible in the background. It consists of several circular nodes connected by lines. The nodes are colored in shades of blue, green, yellow, and orange. The lines represent the network connections between the nodes, forming a complex, interconnected web.

A Função do NameNode no Processo de
Gravação no HDFS



Workflow do Cluster Hadoop

NameNode

O NameNode é o controlador principal do HDFS, que mantém os metadados de todo o sistema de arquivos para o cluster.

Principais características do NameNode:

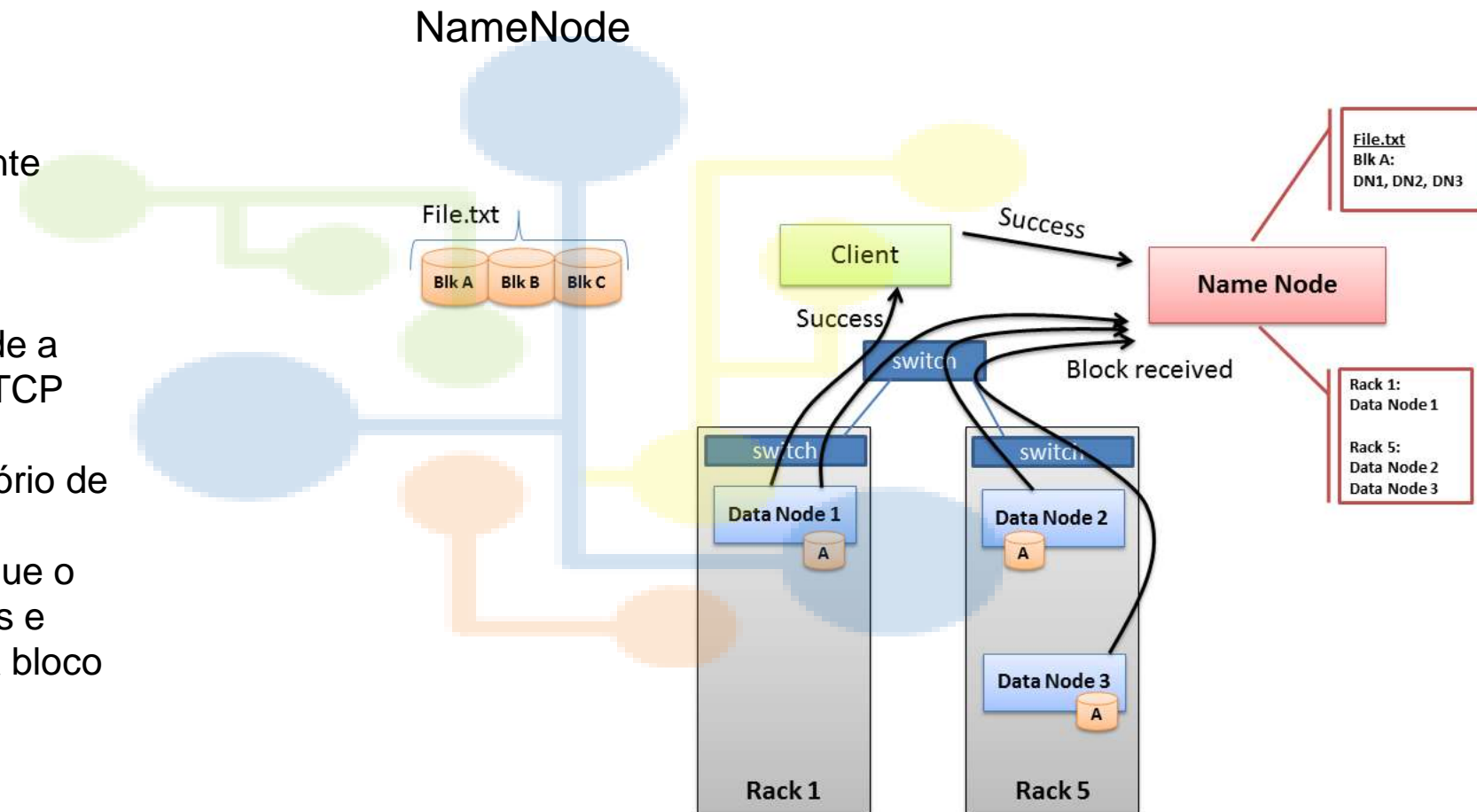
- ☐ Mantém o track de como cada bloco compõe um arquivo e a localização de cada bloco no cluster
- ☐ O NameNode não contém qualquer bloco de dados
- ☐ Direciona o cliente para os DataNodes e mantém o histórico de condições de cada DataNode
- ☐ Garante que cada bloco de dado atende aos critérios mínimos definidos pela política de replicação



Workflow do Cluster Hadoop

O NameNode funciona da seguinte forma:

- ✓ Os DataNodes enviam sinais (heartbeats) para o NameNode a cada 3 segundos através de TCP Handshake.
- ✓ Cada décimo sinal é um relatório de bloco.
- ✓ O relatório de bloco permite que o NameNode crie os metadados e garanta que 3 cópias de cada bloco existam em nodes diferentes.

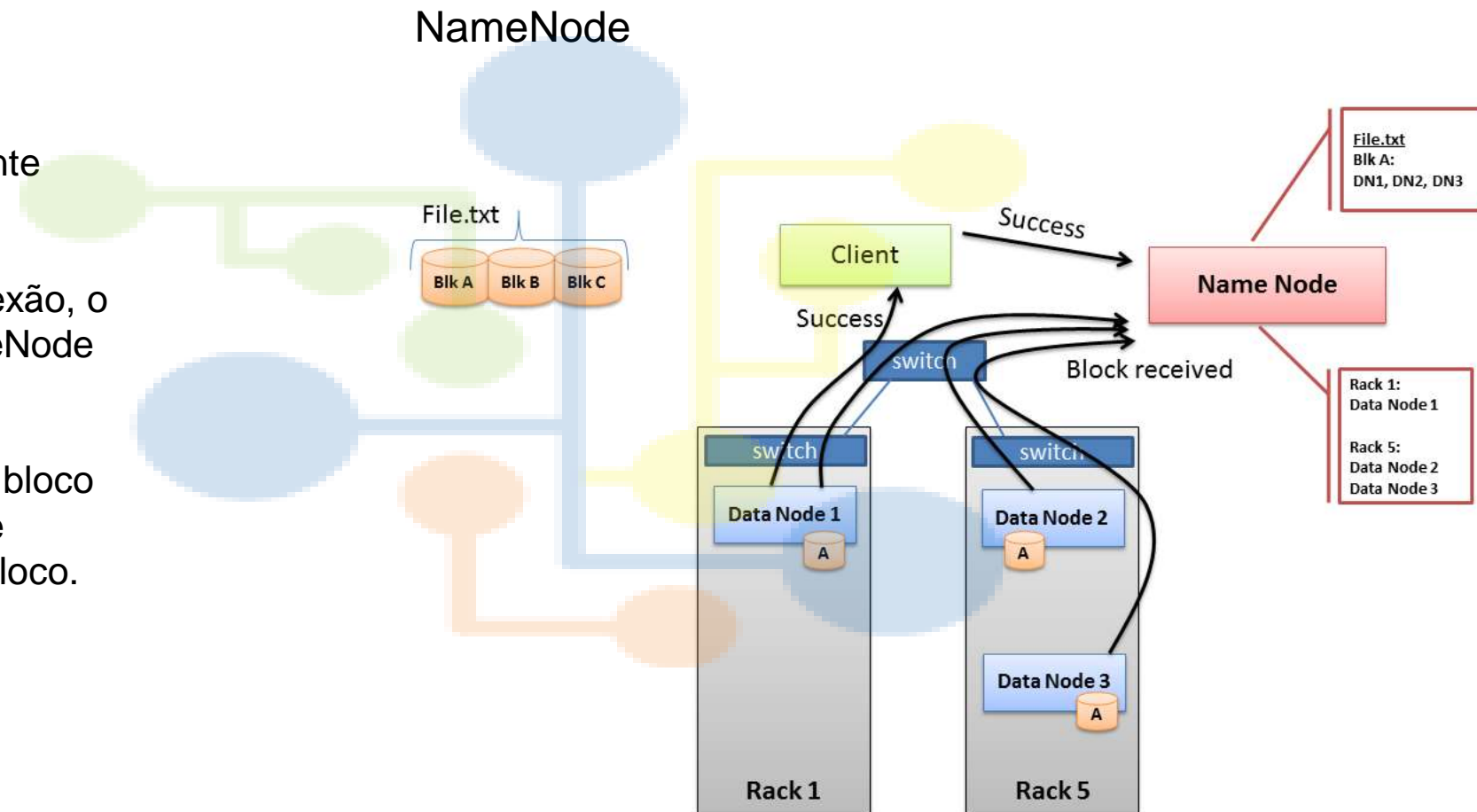




Workflow do Cluster Hadoop


O NameNode funciona da seguinte forma:

- ✓ Se o DataNode fica sem conexão, o sinal não é enviado e o NameNode deixa de considerar aquele DataNode
- ✓ O NameNode então replica o bloco para outro DataNode, sempre mantendo 3 cópias de cada bloco.





Workflow do Cluster Hadoop

A faint, stylized diagram in the background shows a central blue node connected to several other nodes of different colors (green, yellow, orange, and blue) via lines, representing a data flow or cluster topology.

Workflow de Leitura de
Dados no HDFS

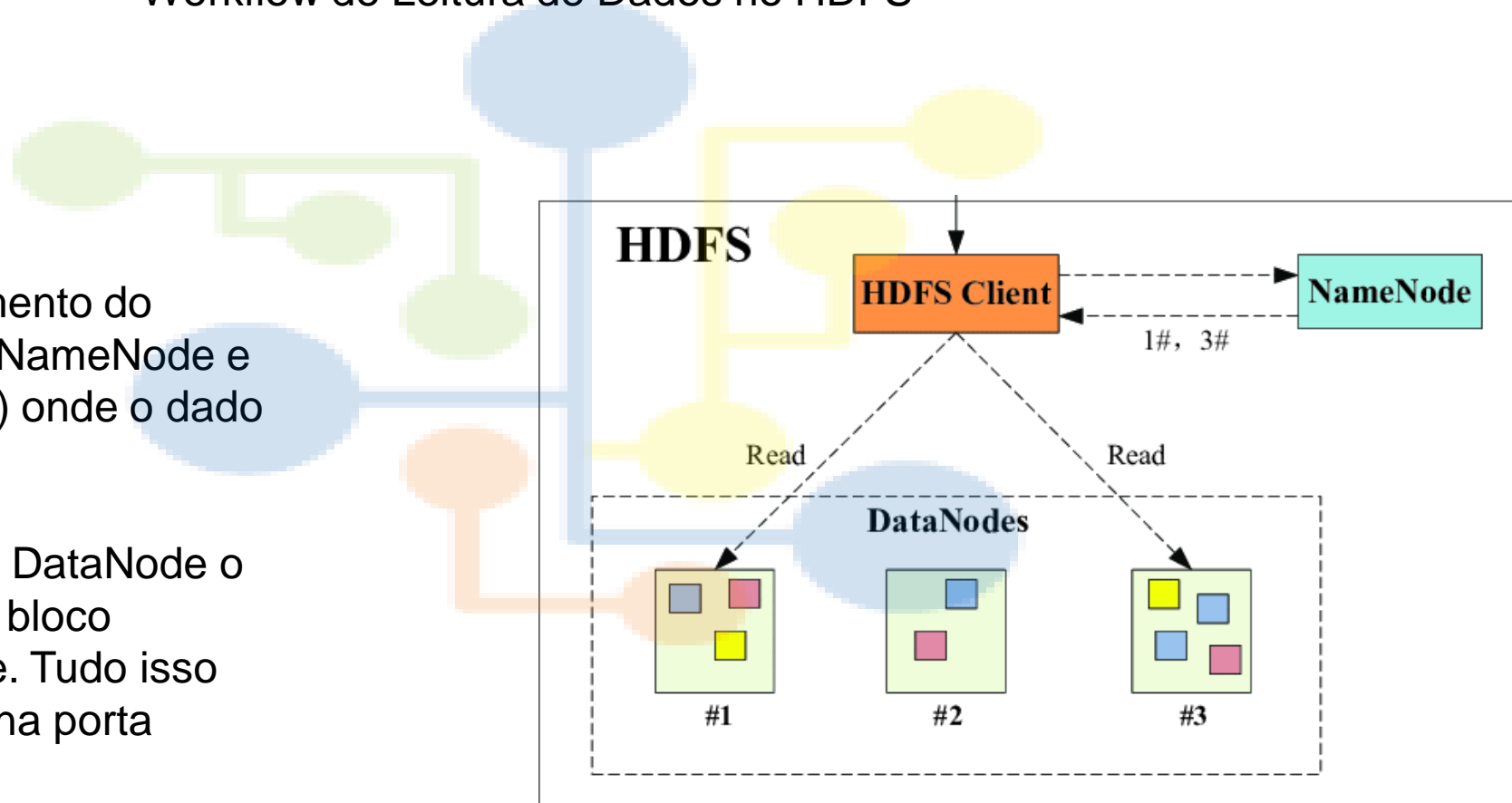


Workflow do Cluster Hadoop

Workflow de Leitura de Dados no HDFS

Leitura dos dados do HDFS:

- Para recuperar um documento do HDFS, o cliente aciona o NameNode e solicita o endereço (bloco) onde o dado está armazenado.
- O cliente então solicita ao DataNode o dado, com o endereço do bloco fornecido pelo NameNode. Tudo isso ocorre via protocolo TCP na porta 50010.





Planejamento do Cluster Hadoop





Planejamento do Cluster Hadoop

Fatores para Planejamento do Cluster Hadoop





Planejamento do Cluster Hadoop

Fatores para Planejamento do Cluster Hadoop

Objetivo

Volume de dados x Alta disponibilidade

Serviços

MapReduce (JobTracker, TaskTracker),
HDFS (NameNode, DataNode), Storage (NFS, SAN)

Layout

Pseudo-Distribuído para desenvolvimento e Totalmente Distribuído para produção
Local / Nuvem



Hardware e Configuração de Rede do Cluster Hadoop

A faded background diagram showing a network topology with several nodes (blue, green, yellow, orange) connected by lines, representing a cluster configuration.



Hardware e Configuração de Rede do Cluster Hadoop

Hardware e Configuração de Rede do Cluster Hadoop

Worker

Configuração	Descrição
Storage	Em um ambiente de intensivo i/o, recomenda-se 12 discos SATA 7200 RPM de 2 TB cada um, para balanceamento entre custo e performance. RAID não é recomendado em máquinas com serviços workers do Hadoop.
Memória	Nodes slaves requerem normalmente entre 24 e 48 GB de memória RAM. Memória não utilizada será consumida por outras aplicações Hadoop.
Processador	Processadores com clock médio e menos de 2 sockets são recomendados.
Rede	Cluster de tamanho considerável, tipicamente requer links de 1 GB para todos os nodes em um rack com 20 nodes.



Hardware e Configuração de Rede do Cluster Hadoop

Hardware e Configuração de Rede do Cluster Hadoop

Master

Configuração	Descrição
Storage	Deve-se utilizar 2 servidores: um para o NameNode Principal e outro para o Secundário. O Master deve ter pelo menos 4 volumes de storage redundantes, seja local ou em rede.
Memória	64 GB de RAM suportam aproximadamente 100 milhões de arquivos.
Processador	16 ou 24 CPU's para suportar o tráfego de mensagens.



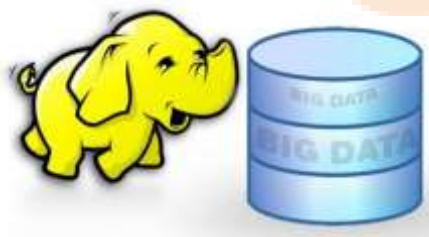
Hardware e Configuração de Rede do Cluster Hadoop

Estas são apenas recomendações e que podem variar de acordo com os fatores para o planejamento do cluster: objetivo, serviços e layout.



Hardware e Configuração de Rede do Cluster Hadoop

Instalação do Hadoop





Hardware e Configuração de Rede do Cluster Hadoop

Single Node x Multi Node

Cluster Single Node	Cluster Multi Node
Hadoop é instalado em um único servidor (node)	Hadoop é instalado em diversos nodes (entre algumas dezenas, até milhares)
Clusters Single Node são usados para processos triviais e operações simples de MapReduce e HDFS. Pode ser usado em ambiente de testes.	Clusters Multi Node são usados para computação complexa, incluindo processamento analítico.



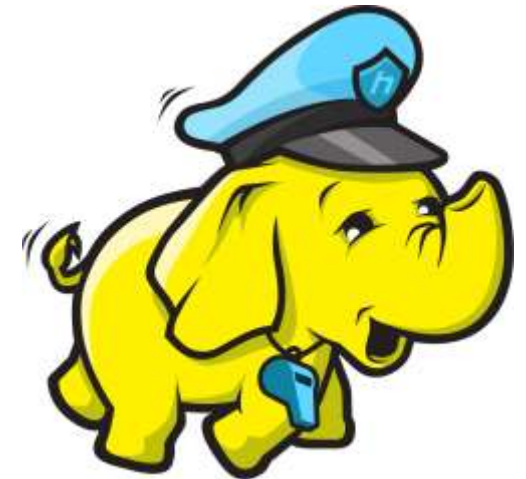
Quando Usar e Quando Não Usar o HDFS?





Quando Usar e Quando Não Usar o HDFS?

- Hadoop Distributed File System (HDFS) é um framework distribuído e extremamente tolerante a falha.
- Foi concebido para processar grandes volumes de dados.
- O conceito do HDFS é baseado no Unix.
- O HDFS é similar a outros frameworks de arquivos distribuídos, mas com algumas diferenças:
 - O HDFS possui um modelo chamado "write-once-read-many-times" (WORM), que significa: escreva uma vez e leia quantas vezes quiser.
 - Eficiente controle de concorrência.
 - Redireciona atividades (jobs) em caso de falhas.





Quando Usar e Quando Não Usar o HDFS?

Quando usar o
HDFS?

Grande quantidade de dados a serem armazenados

Streams de dados constantes que requerem acesso

Apenas equipamentos simples estão disponíveis



Quando Usar e Quando Não Usar o HDFS?

Quando NÃO usar
o HDFS?

Quantidade considerável de arquivos pequenos

Composições variadas (muitos arquivos em formatos diferentes)

Acesso de baixa latência aos dados



Obrigado
