



Engenharia de Dados com Hadoop e Spark



Bem-vindo(a)





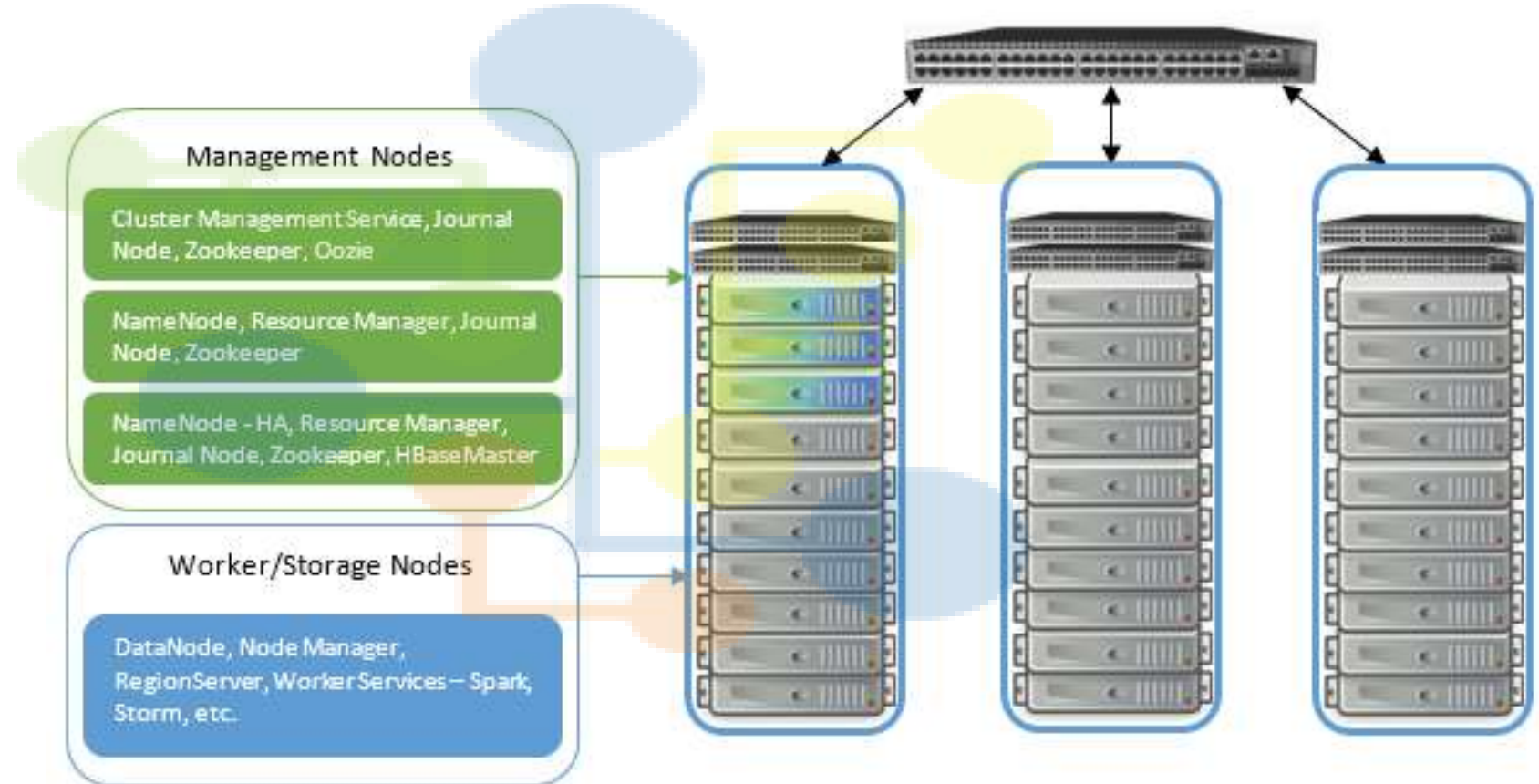
Administração e Manutenção do Hadoop





Administração e Manutenção do Hadoop

Cluster Hadoop





Desafios na Administração e Manutenção de um Cluster Hadoop





Administração e Manutenção do Hadoop

Desafios na gestão do cluster Hadoop

Falta de gestão de configuração

Baixa alocação de recursos

Gargalos de rede

Falta de métricas de monitoramento





Administração e Manutenção do Hadoop

Desafios na gestão do cluster Hadoop

Medidas drásticas
para resolver
problemas
simples

Pontos únicos de
falha

Utilização dos
valores default
para os
parâmetros

Falta de
profissionais
qualificados





Administração e Manutenção do Hadoop



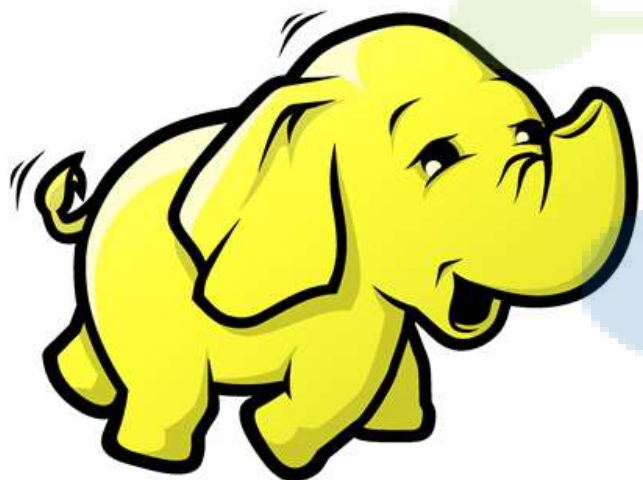


NameNode e Estrutura de Diretórios

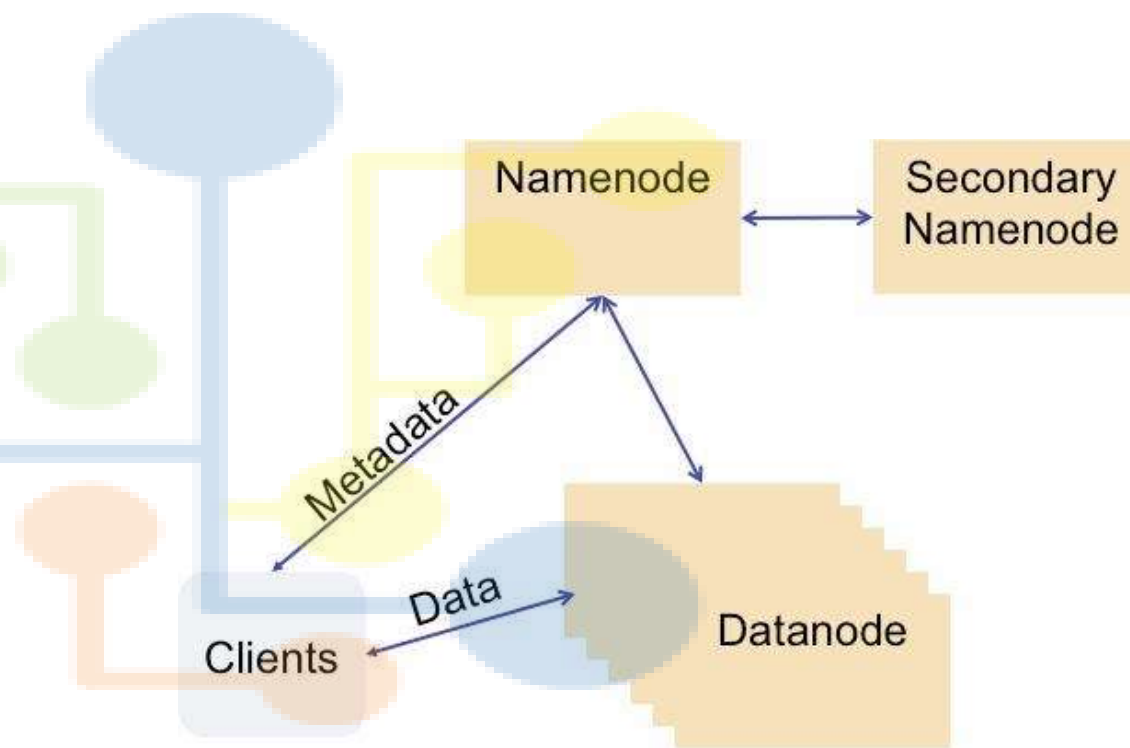




Estrutura de Diretórios do NameNode

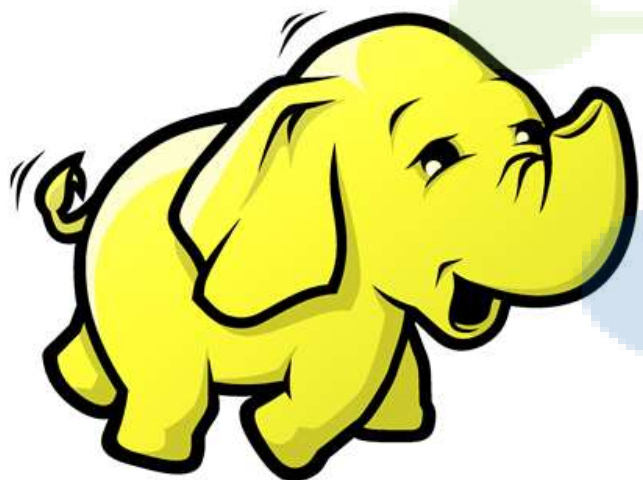


NAMENODE

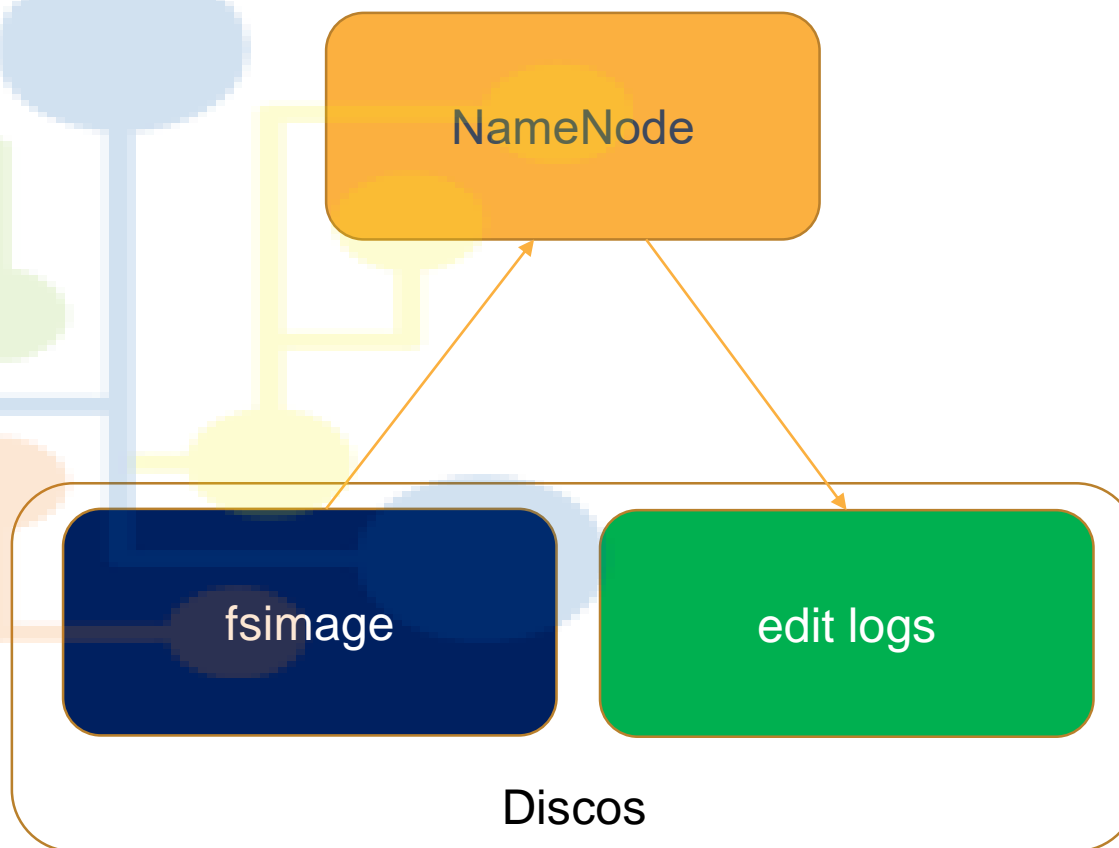




Estrutura de Diretórios do NameNode



NAMENODE





Estrutura de Diretórios do NameNode

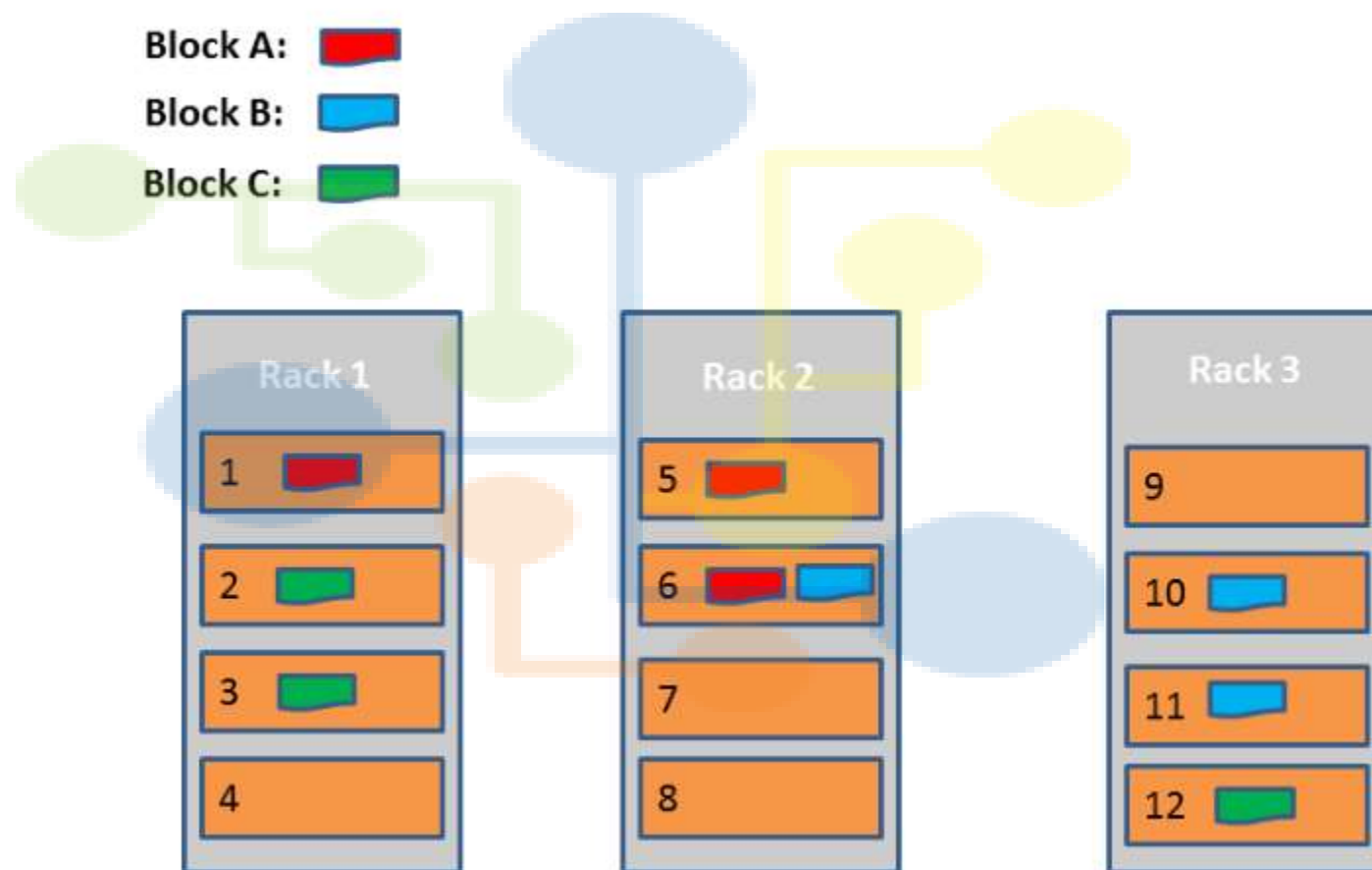
Com o passar do tempo, o número de arquivos edit-log pode se tornar grande demais, sendo necessária uma atualização do fs-image.

Essa é a função do SecondaryNameNode, que veremos mais adiante.



Estrutura de Diretórios do NameNode

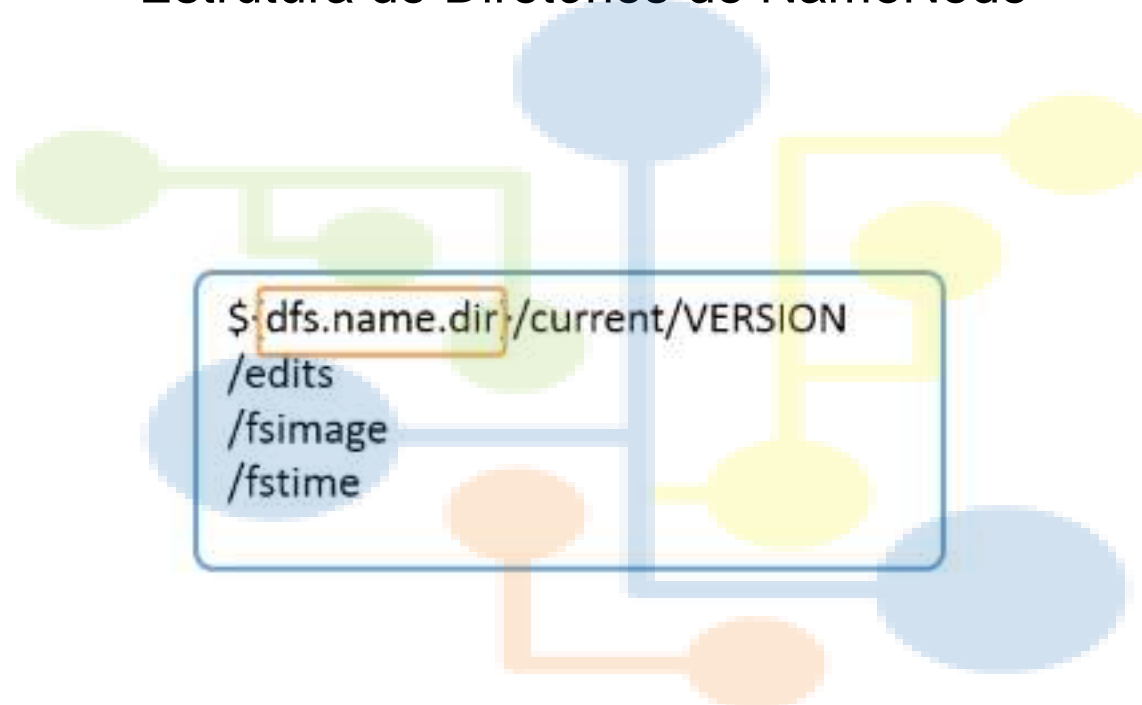
Rack Awareness





Estrutura de Diretórios do NameNode

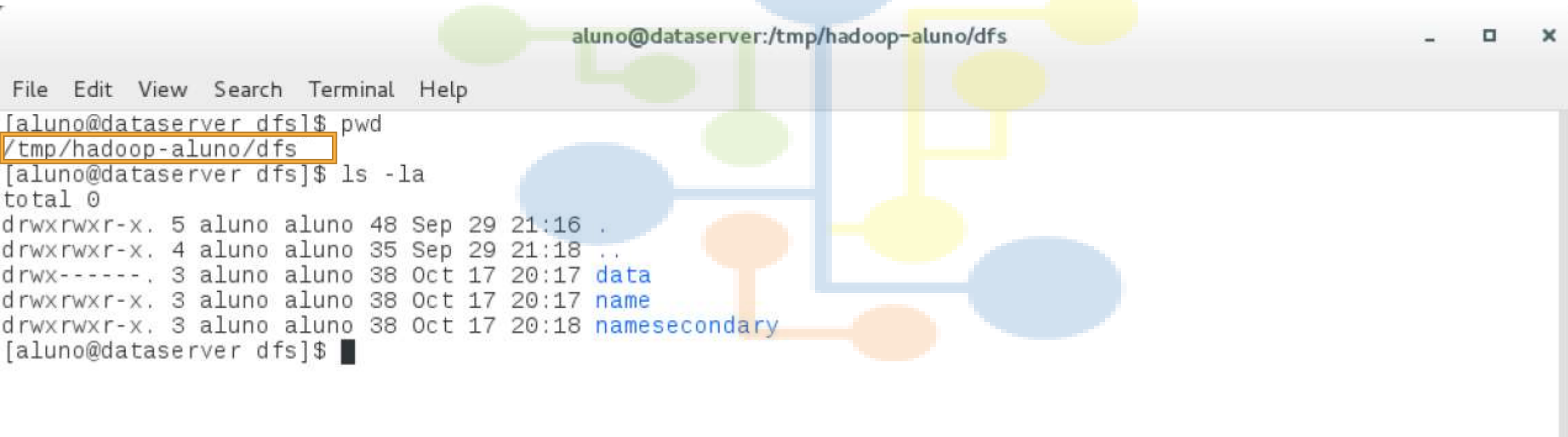
Estrutura de Diretórios do NameNode





Estrutura de Diretórios do NameNode

Estrutura de Diretórios do NameNode



A terminal window titled 'aluno@dataserver:tmp/hadoop-aluno/dfs' displays the output of the 'ls -la' command. The output shows the directory structure of the NameNode, including the 'data' directory and the 'name' directory. The 'name' directory is highlighted with a yellow box. The 'name' directory contains a subdirectory named 'namesecondary'.

```
File Edit View Search Terminal Help
[aluno@dataserver dfs]$ pwd
/tmp/hadoop-aluno/dfs
[aluno@dataserver dfs]$ ls -la
total 0
drwxrwxr-x. 5 aluno aluno 48 Sep 29 21:16 .
drwxrwxr-x. 4 aluno aluno 35 Sep 29 21:18 ..
drwx----- 3 aluno aluno 38 Oct 17 20:17 data
drwxrwxr-x. 3 aluno aluno 38 Oct 17 20:17 name
drwxrwxr-x. 3 aluno aluno 38 Oct 17 20:18 namesecondary
[aluno@dataserver dfs]$
```



Estrutura de Diretórios do NameNode

aluno@dataserver:/tmp/hadoop-aluno/dfs/name/current

File Edit View Search Terminal Help

[aluno@dataserver current]\$ pwd

/tmp/hadoop-aluno/dfs/name/current

[aluno@dataserver current]\$ ls -la

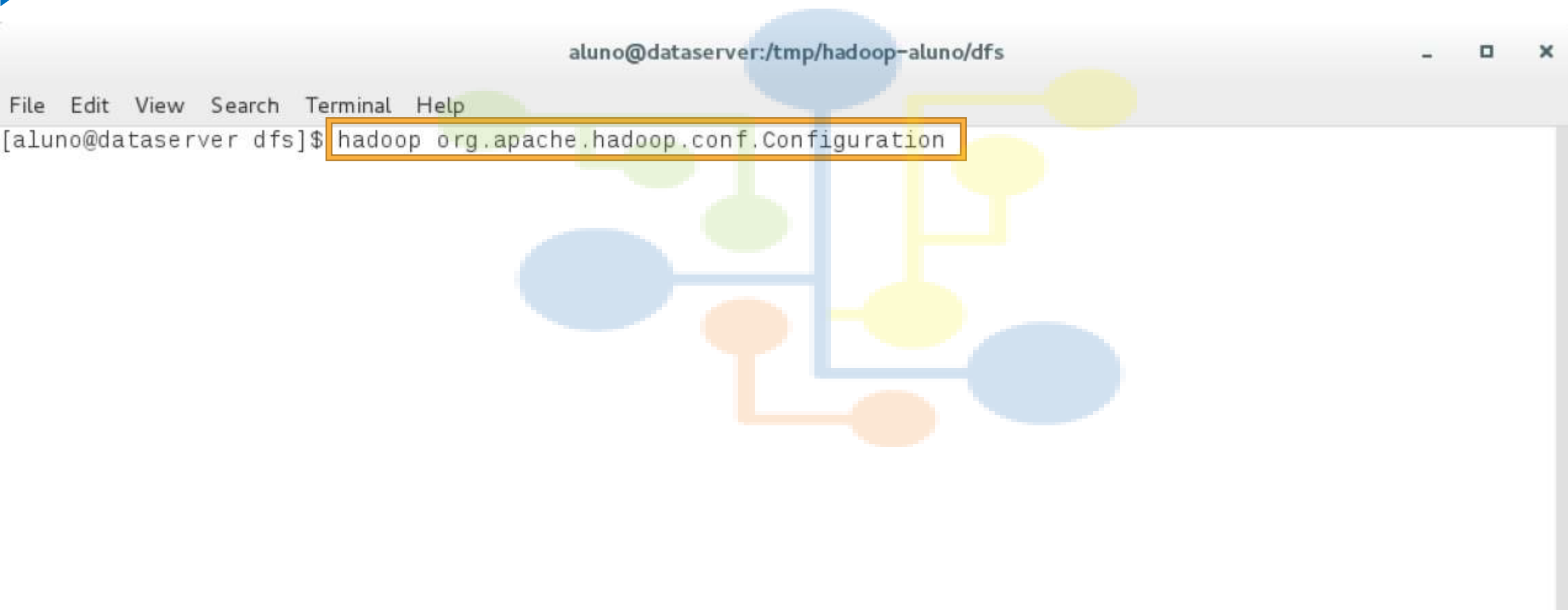
total 7244

drwxrwxr-x.	2	aluno	aluno	4096	Oct 17 20:28	.
drwxrwxr-x.	3	aluno	aluno	38	Oct 17 20:17	..
-rw-rw-r--.	1	aluno	aluno	1048576	Sep 30 19:45	edits_00000000000000000000120-000000000000000000120
-rw-rw-r--.	1	aluno	aluno	1048576	Oct 16 18:30	edits_00000000000000000000121-000000000000000000121
-rw-rw-r--.	1	aluno	aluno	1048576	Oct 16 18:42	edits_00000000000000000000122-000000000000000000122
-rw-rw-r--.	1	aluno	aluno	1048576	Oct 16 18:48	edits_00000000000000000000123-000000000000000000123
-rw-rw-r--.	1	aluno	aluno	1048576	Oct 16 19:02	edits_00000000000000000000124-000000000000000000124
-rw-rw-r--.	1	aluno	aluno	42	Oct 16 20:06	edits_00000000000000000000125-000000000000000000126
-rw-rw-r--.	1	aluno	aluno	38394	Oct 16 21:06	edits_00000000000000000000127-000000000000000000441
-rw-rw-r--.	1	aluno	aluno	42	Oct 16 22:06	edits_00000000000000000000442-000000000000000000443
-rw-rw-r--.	1	aluno	aluno	1048576	Oct 16 22:06	edits_00000000000000000000444-000000000000000000444
-rw-rw-r--.	1	aluno	aluno	1048576	Oct 17 20:17	edits_inprogress_00000000000000000000445
-rw-rw-r--.	1	aluno	aluno	2587	Oct 16 22:06	fsimage_00000000000000000000443
-rw-rw-r--.	1	aluno	aluno	62	Oct 16 22:06	fsimage_00000000000000000000443.md5
-rw-rw-r--.	1	aluno	aluno	2587	Oct 17 20:17	fsimage_00000000000000000000444
-rw-rw-r--.	1	aluno	aluno	62	Oct 17 20:17	fsimage_00000000000000000000444.md5
-rw-rw-r--.	1	aluno	aluno	4	Oct 17 20:17	seen_txid
-rw-rw-r--.	1	aluno	aluno	201	Oct 17 20:17	VERSION

[aluno@dataserver current]\$



Estrutura de Diretórios do NameNode





Estrutura de Diretórios do NameNode

aluno@dataserver:~

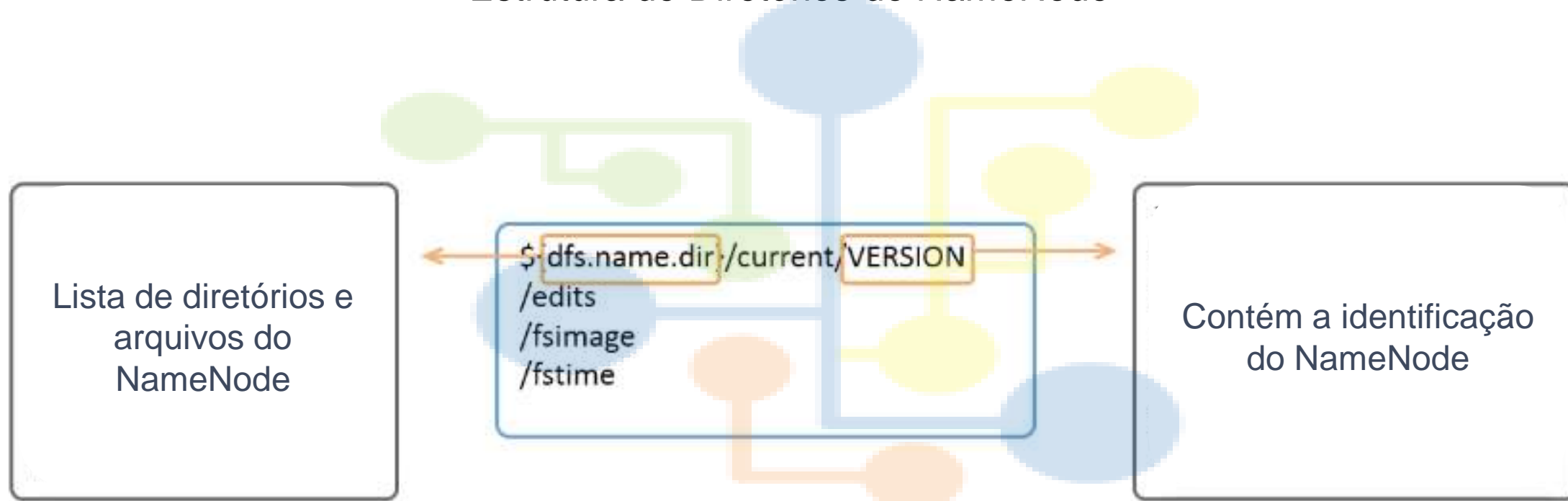
File Edit View Search Terminal Help

```
<property><name>ipc.client.connection.maxidletime</name><value>10000</value><source>core-default.xml</source></property>
<property><name>ipc.client.connect.timeout</name><value>20000</value><source>core-default.xml</source></property>
<property><name>hadoop.security.uid.cache.secs</name><value>14400</value><source>core-default.xml</source></property>
<property><name>ipc.client.ping</name><value>true</value><source>core-default.xml</source></property>
<property><name>ipc.client.kill.max</name><value>10</value><source>core-default.xml</source></property>
<property><name>ipc.client.connect.max.retries</name><value>10</value><source>core-default.xml</source></property>
<property><name>ipc.ping.interval</name><value>60000</value><source>core-default.xml</source></property>
<property><name>io.seqfile.local.dir</name><value>${hadoop.tmp.dir}/io/local</value><source>core-default.xml</source></property>
<property><name>hadoop.security.crypto.buffer.size</name><value>8192</value><source>core-default.xml</source></property>
<property><name>io.native.lib.available</name><value>true</value><source>core-default.xml</source></property>
<property><name>io.file.buffer.size</name><value>4096</value><source>core-default.xml</source></property>
<property><name>io.serializations</name><value>org.apache.hadoop.io.serializer.WritableSerialization,org.apache.hadoop.io.serializer.avro.AvroSpecificSerialization,org.apache.hadoop.io.serializer.avro.AvroReflectSerialization</value><source>core-default.xml</source></property>
<property><name>tfile.fs.input.buffer.size</name><value>262144</value><source>core-default.xml</source></property>
<property><name>hadoop.registry.zk.session.timeout.ms</name><value>60000</value><source>core-default.xml</source></property>
<property><name>hadoop.security.group.mapping.ldap.ssl</name><value>false</value><source>core-default.xml</source></property>
<property><name>fs.df.interval</name><value>60000</value><source>core-default.xml</source></property>
<property><name>hadoop.http.authentication.kerberos.keytab</name><value>${user.home}/hadoop.keytab</value><source>core-default.xml</source></property>
<property><name>s3native.client.write.packet.size</name><value>65536</value><source>core-default.xml</source></property>
<property><name>s3native.replication</name><value>3</value><source>core-default.xml</source></property>
<property><name>hadoop.http.cross-origin.allowed-headers</name><value>X-Requested-With,Content-Type,Accept,Origin</value><source>core-default.xml</source></property>
<property><name>tfile.io.chunk.size</name><value>1048576</value><source>core-default.xml</source></property>
<property><name>hadoop.ssl.hostname.verifier</name><value>DEFAULT</value><source>core-default.xml</source></property>
[aluno@dataserver ~]$
```



Estrutura de Diretórios do NameNode

Estrutura de Diretórios do NameNode





Estrutura de Diretórios do NameNode

Poucos arquivos grandes consomem menos memória



A Importância do Secondary NameNode



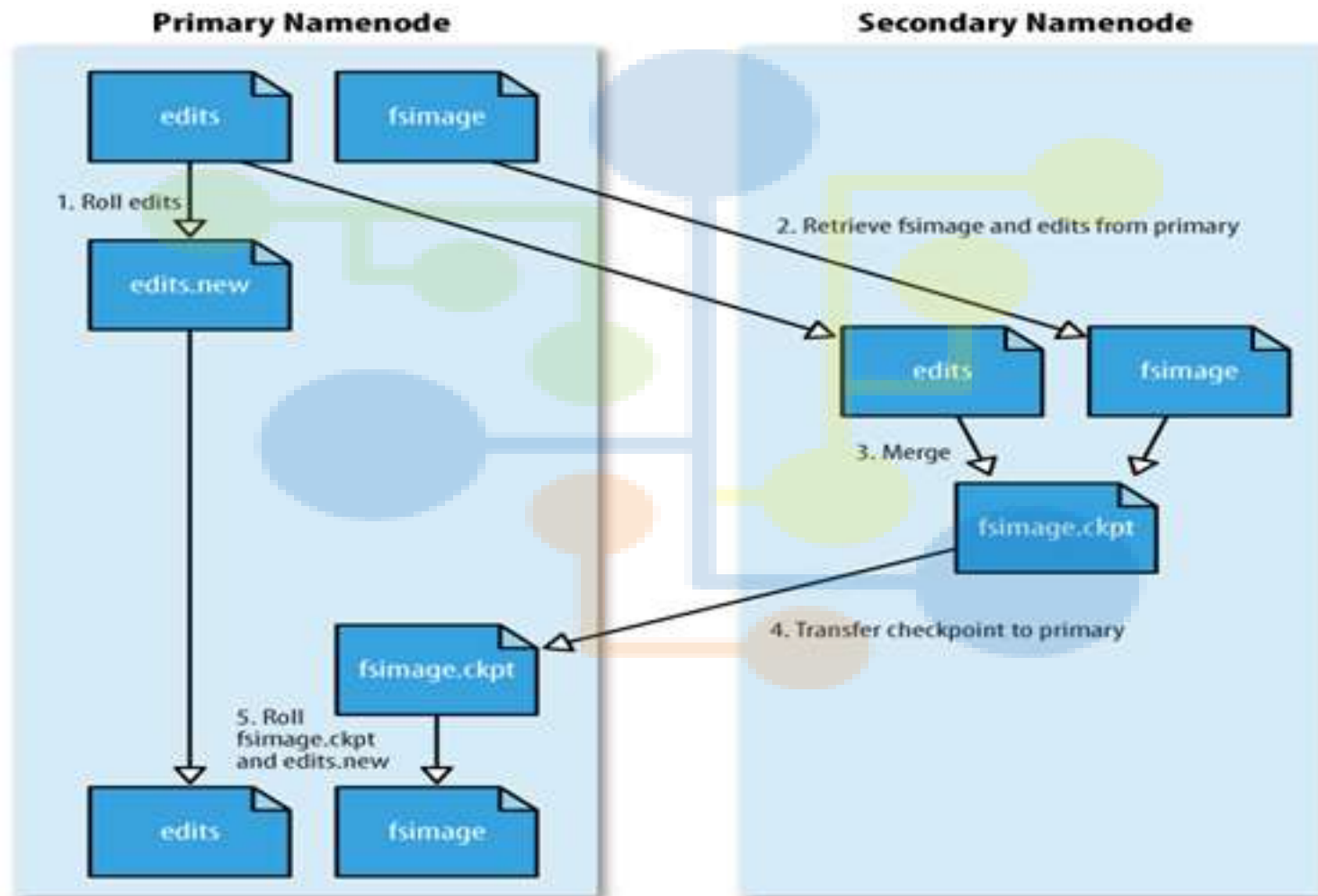


Secondary NameNode

O Secondary NameNode é o processo responsável por sincronizar os arquivos edit logs com a imagem do fsimage, para gerar um novo fsimage mais atualizado.



Secondary NameNode





Secondary NameNode

O principal objetivo do Secondary NameNode é ajudar a reconstruir o NameNode no caso desse vir a falhar!



DataNodes e Estrutura de Diretórios





Estrutura de Diretórios do DataNode

Os DataNodes não precisam ser formatados (como o NameNode), uma vez que eles criam seus diretórios no storage automaticamente na inicialização.



Estrutura de Diretórios do DataNode

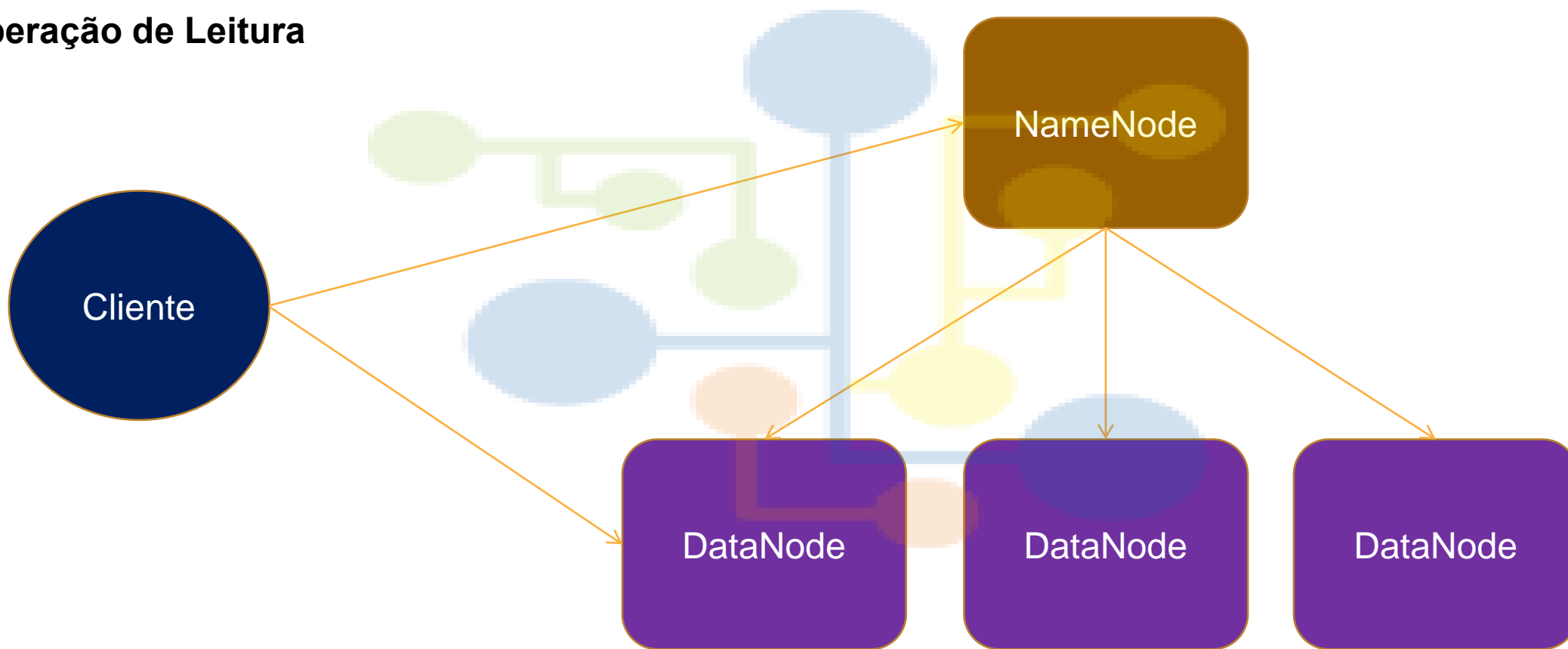
A estrutura do DataNode é muito similar a do NameNode, com um arquivo VERSION com informações sobre o servidor e os arquivos binários de operação do serviço DataNode. Esse diretório é definido pelo parâmetro `dfs.data.node.dir`, nos arquivos de configuração do Hadoop.

```
hadoop@dataserver:/opt/hadoop/dfs/data/current
File Edit View Search Terminal Help
(base) [hadoop@dataserver ~]$ cd /opt/hadoop/dfs/data/
(base) [hadoop@dataserver data]$ ls -la
total 0
drwx-----. 3 hadoop hadoop 21 Oct 20 13:06 .
drwxrwxr-x. 5 hadoop hadoop 61 Oct 20 12:51 ..
drwxrwxr-x. 3 hadoop hadoop 66 Oct 20 12:01 current
(base) [hadoop@dataserver data]$ cd current/
(base) [hadoop@dataserver current]$ ls
BP-2051156013-127.0.0.1-1563253115551 VERSION
(base) [hadoop@dataserver current]$ ls -la
total 4
drwxrwxr-x. 3 hadoop hadoop 66 Oct 20 12:01 .
drwx-----. 3 hadoop hadoop 21 Oct 20 13:06 ..
drwx-----. 4 hadoop hadoop 54 Oct 20 12:52 BP-2051156013-127.0.0.1-1563253115551
-rw-rw-r--. 1 hadoop hadoop 229 Oct 20 12:52 VERSION
(base) [hadoop@dataserver current]$
```



Estrutura de Diretórios do DataNode

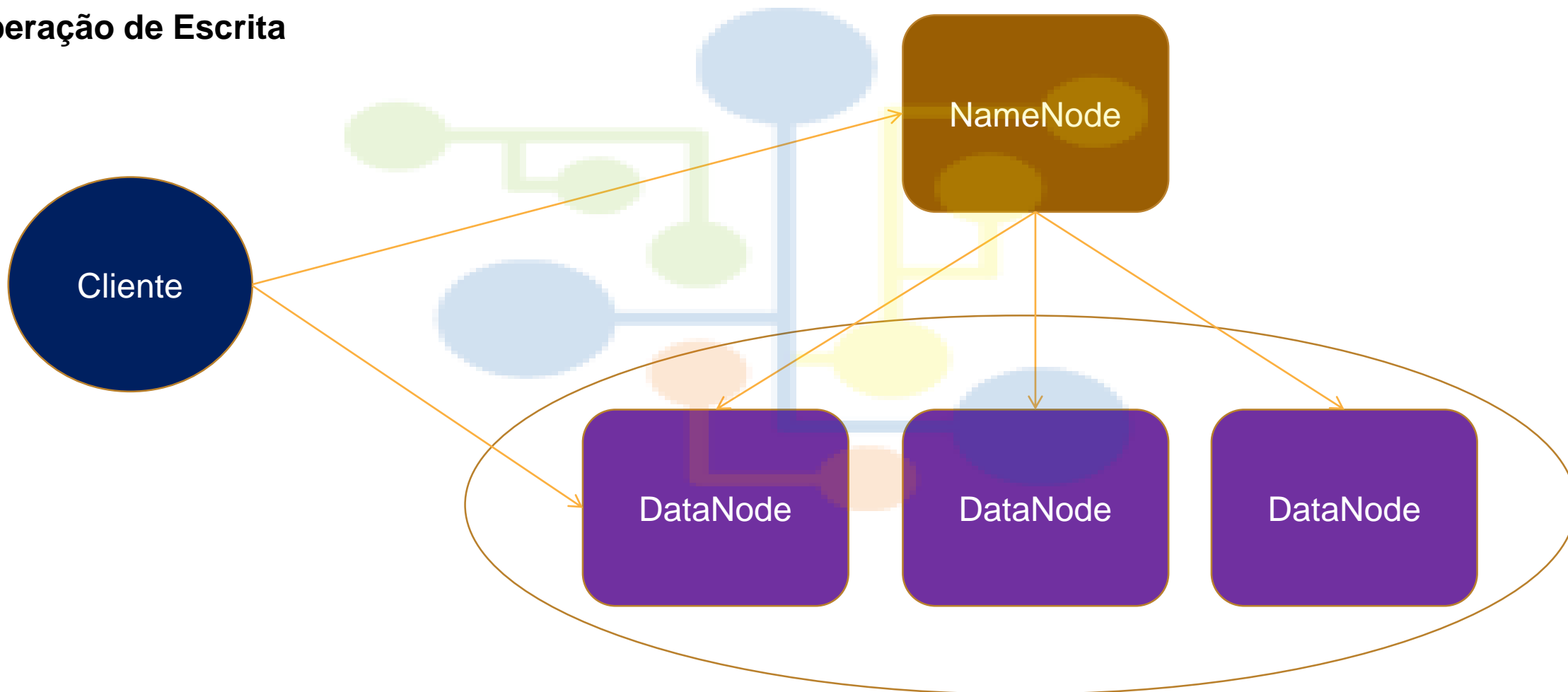
Operação de Leitura





Estrutura de Diretórios do DataNode

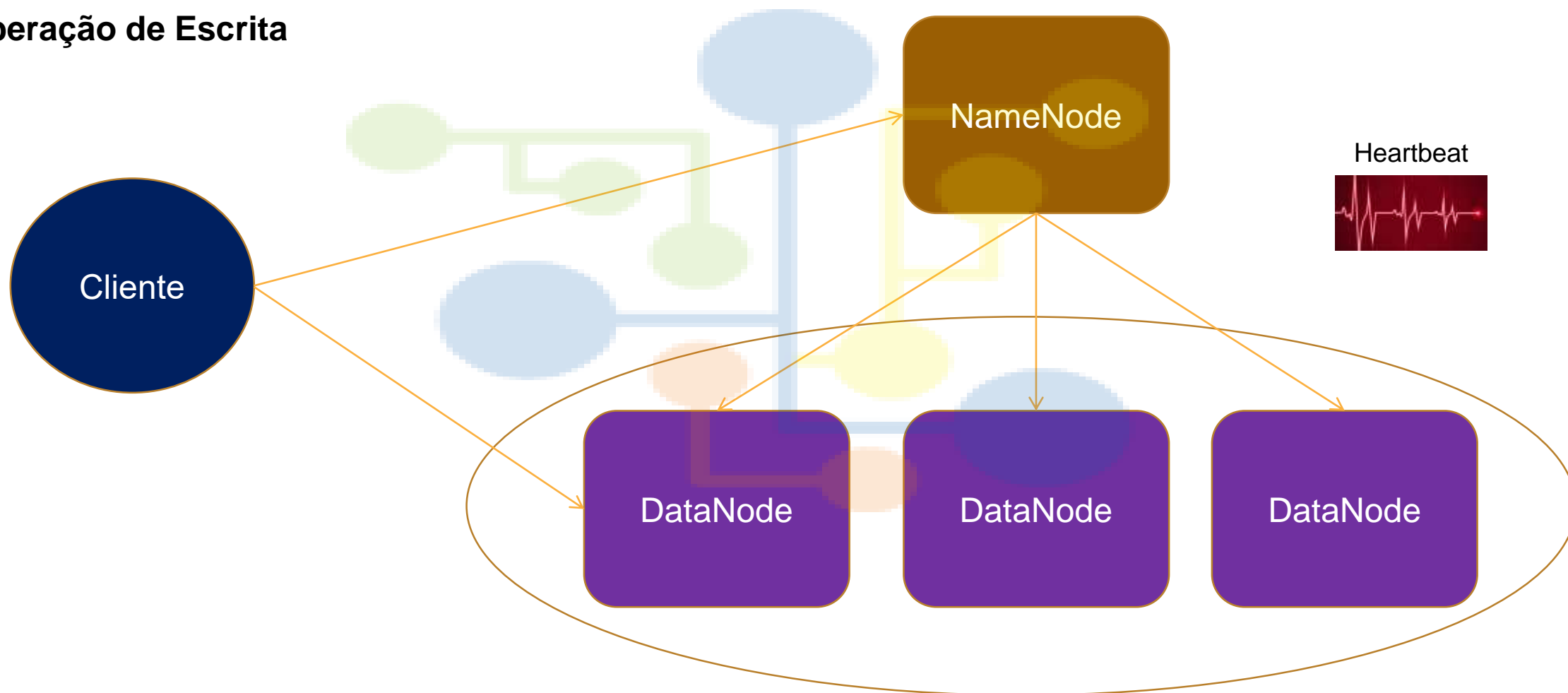
Operação de Escrita





Estrutura de Diretórios do DataNode

Operação de Escrita

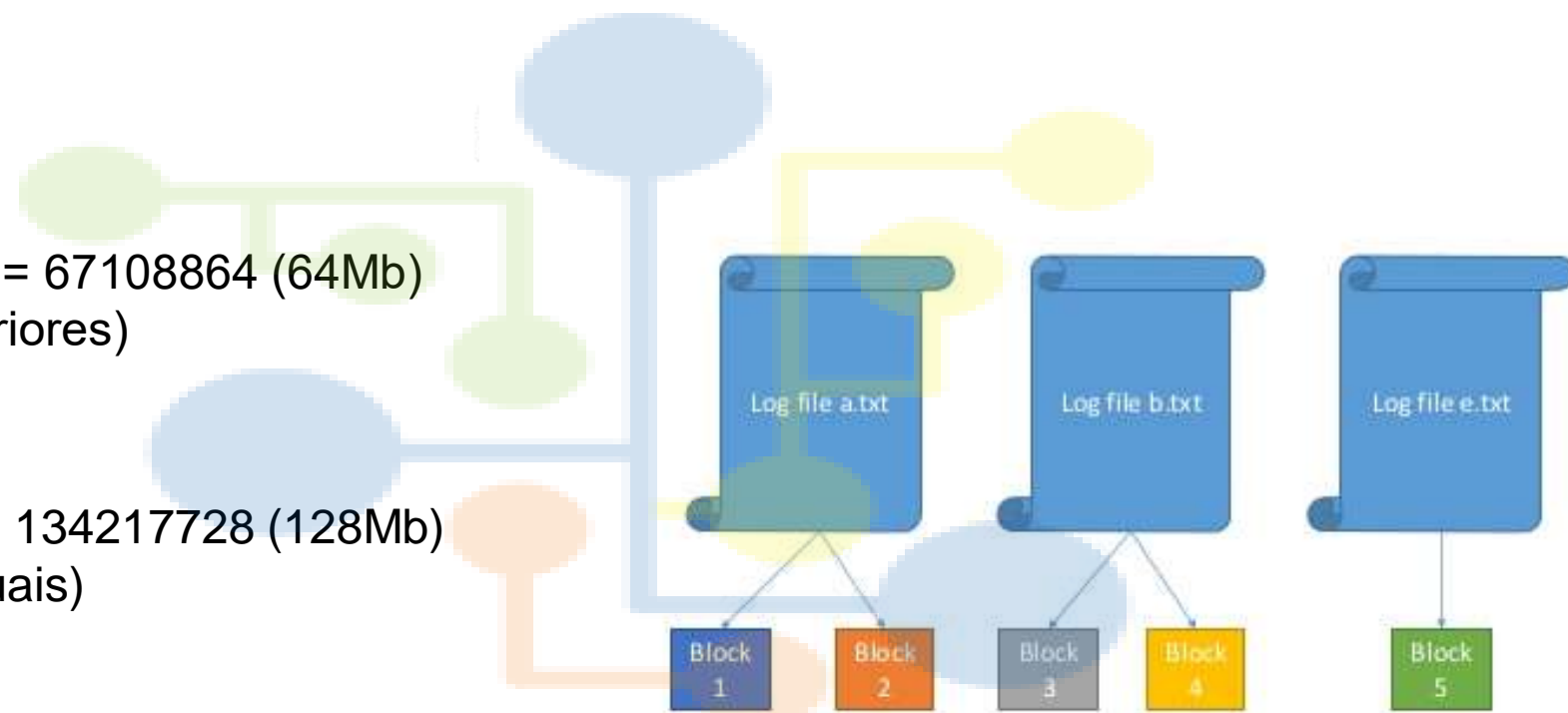




Estrutura de Diretórios do DataNode

Tamanho padrão do bloco = 67108864 (64Mb)
(versão anteriores)

Tamanho padrão do bloco = 134217728 (128Mb)
(versão atuais)





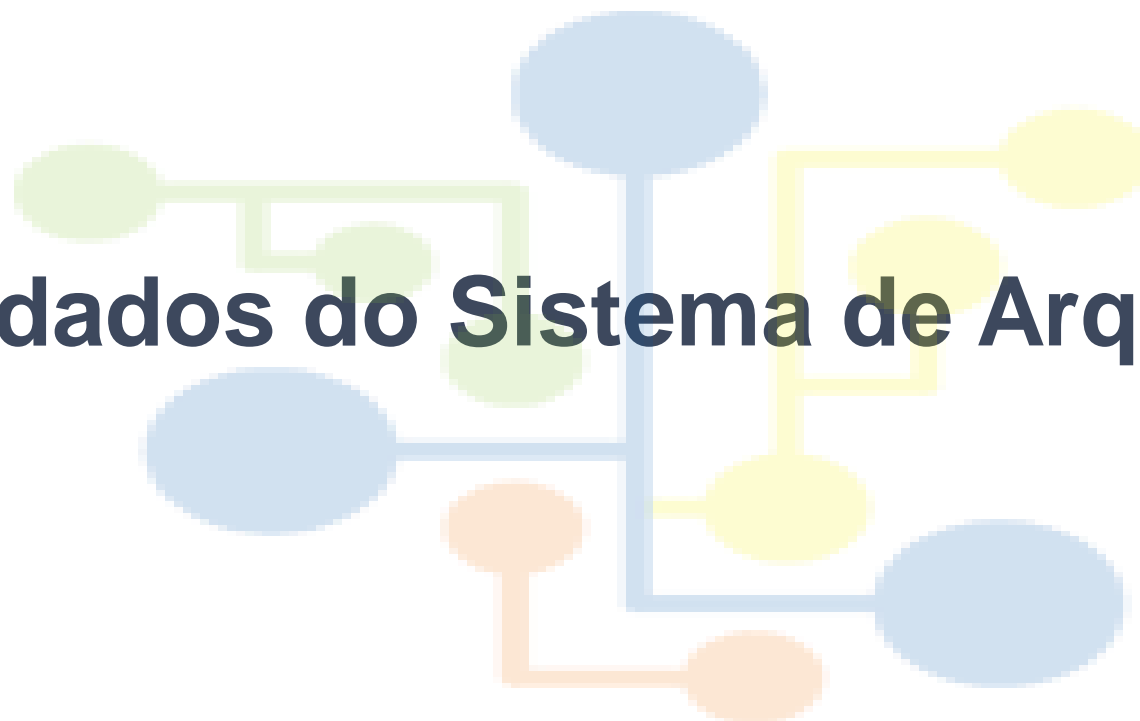
Estrutura de Diretórios do DataNode

Não se usa RAID de discos para fazer cópias de
seguranças dos blocos.





Metadados do Sistema de Arquivos





Metadados do Sistema de Arquivos

Metadados → Informações gerais sobre o cluster e sobre os dados

Dados → Big Data



Metadados do Sistema de Arquivos

ATENÇÃO: Não tente modificar diretórios ou arquivos de metadados. Modificações podem causar interrupção do HDFS ou até mesmo a perda de dados de forma permanente.

O Backup dos metadados é uma tarefa crítica em um cluster Hadoop.



Metadados do Sistema de Arquivos

fsimage	Edit log
Representa uma imagem point-in-time dos metadados do file system	Contém uma série de arquivos, chamados segmentos
O arquivo é sequencial	Os segmentos representam todas as modificações feitas desde a data de criação do fsimage
Pode ser usado para obter o estado mais recente do file system quando o NameNode tiver problemas	Garante que nenhuma operação é perdida devido a uma falha do servidor

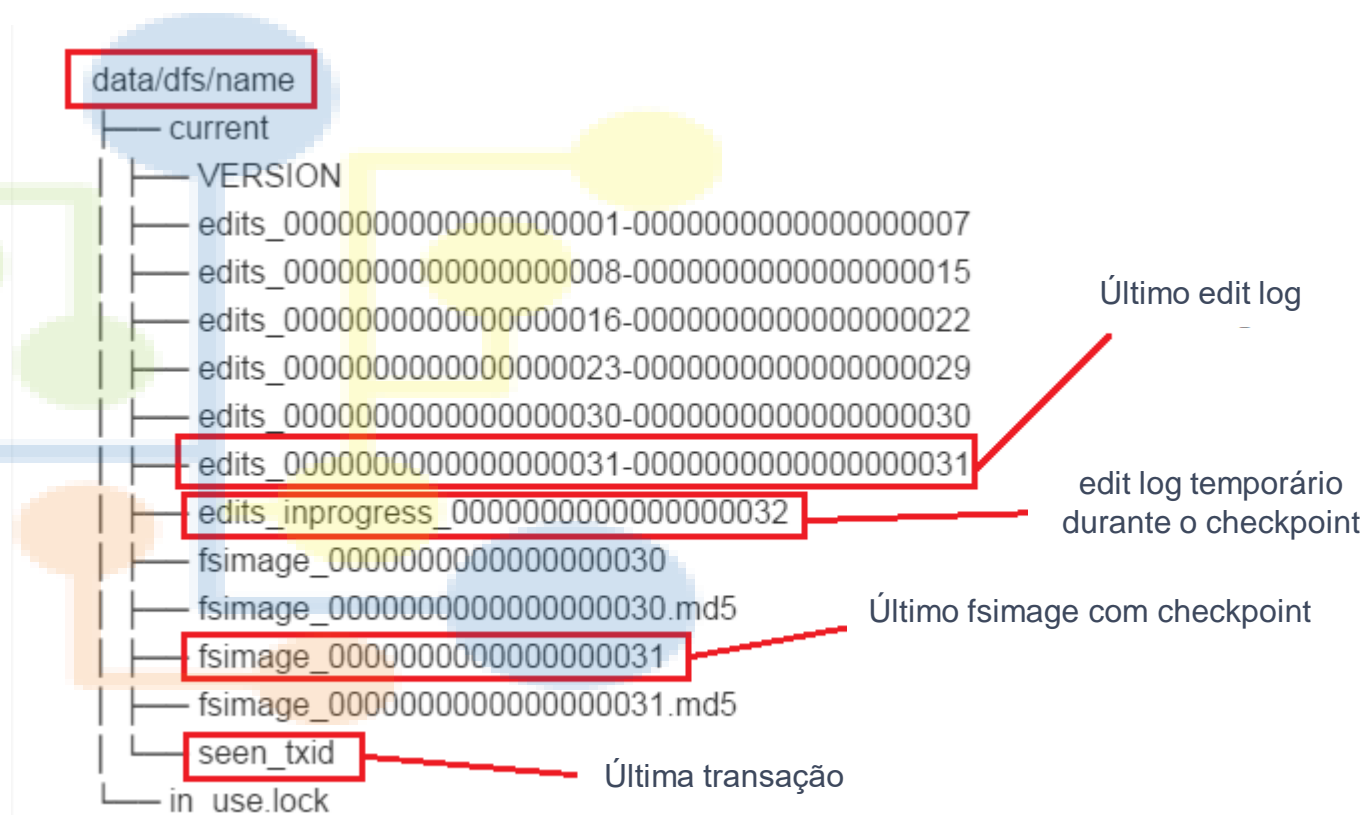
O fsimage não possui qualquer informação sobre os dados armazenados nos DataNodes.



Metadados do Sistema de Arquivos

NameNode

- VERSION
 - Layoutversion
 - namespaceID/clusterID/blockpoolIDstorageType
 - cTime
 - edits_start transaction ID-end transaction ID
 - edits_inprogress__start transaction ID
 - fsimage_end transaction ID
 - seen_txid
- in_use.lock



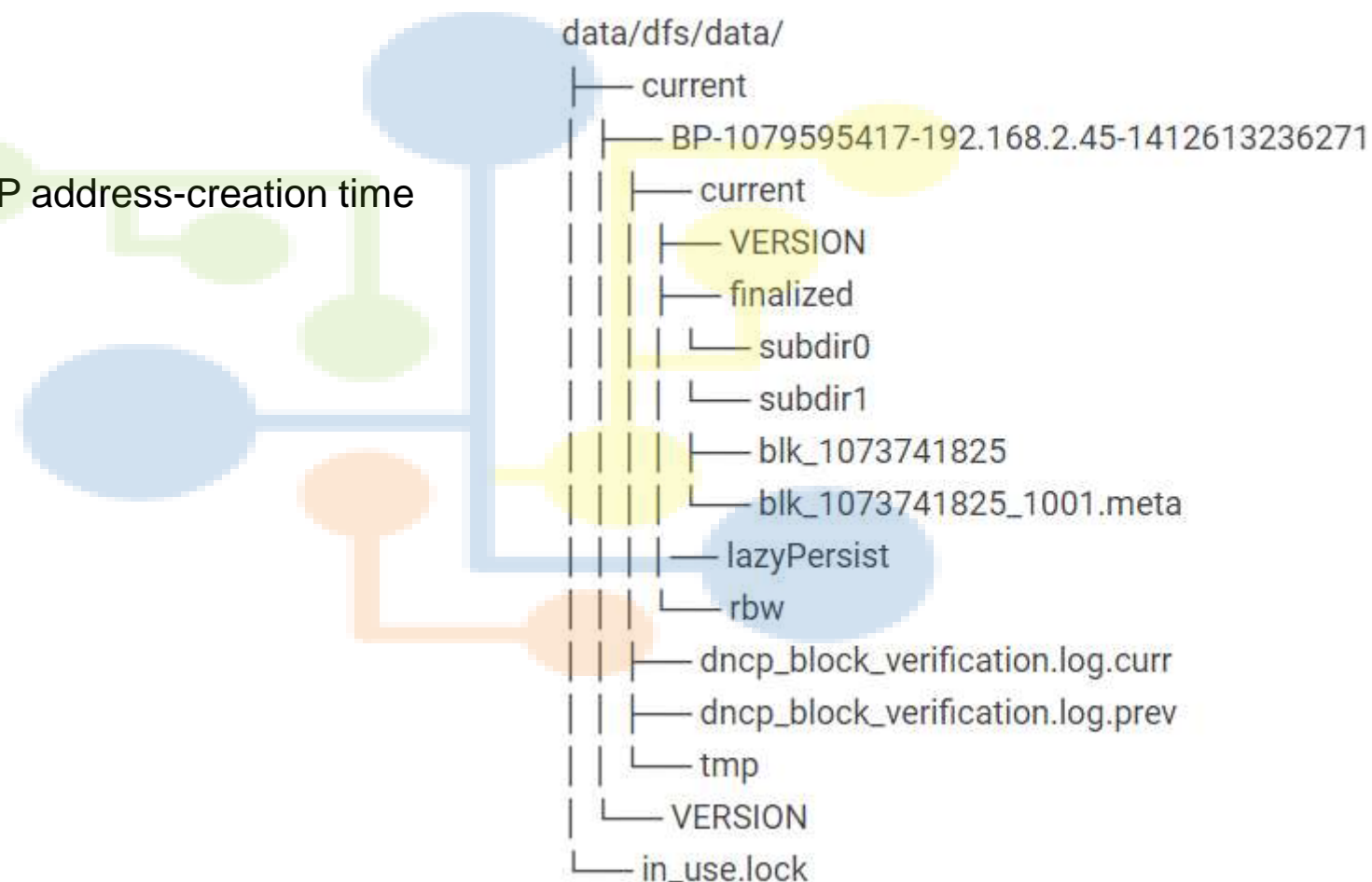
Parâmetro **`dfs.namenode.name.dir`** em `hdfs-site.xml`



Metadados do Sistema de Arquivos

DataNode

- BP-random integer-NameNode-IP address-creation time
- VERSION
 - storageType
 - blockpoolID
 - finalized/rbw
 - lazyPersist
 - dncp_block_verification.log
- in_use.lock





Metadados do Sistema de Arquivos

Parâmetros de Configuração

Lista de Parâmetros para configuração dos diretórios do NameNode e DataNode	
dfs.namenode.name.dir	
dfs.namenode.edits.dir	
dfs.namenode.checkpoint.period	
dfs.namenode.checkpoint.txns	
dfs.namenode.checkpoint.check.period	
dfs.namenode.num.checkpoints.retained	
dfs.namenode.num.extra.edits.retained	
dfs.namenode.edit.log.autoroll.multiplier.threshold	
dfs.namenode.edit.log.autoroll.check.interval.ms	
dfs.datanode.data.dir	



Metadados do Sistema de Arquivos

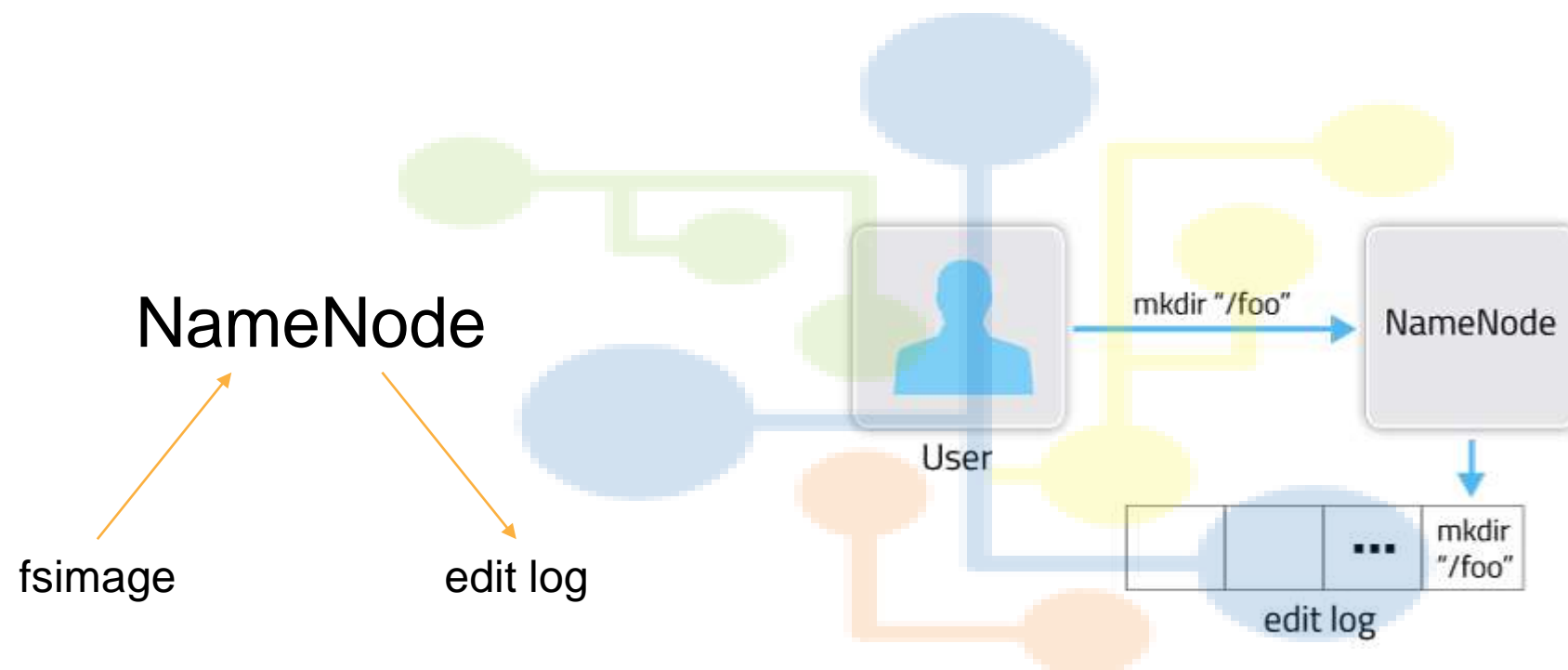
Comandos de Configuração

Comando	Descrição
hdfs namenode	Inicializa o NameNode
hdfs dfsadmin -safemode enter hdfs dfsadmin -saveNamespace	Coloca o NameNode em modo de segurança e realiza um checkpoint
hdfs dfsadmin -rollEdits	Passa de um edit log para outro
hdfs dfsadmin -fetchImage	Obtém a última versão do fsimage (o que pode ser usado para criar um NameNode backup)



Procedimento de Checkpoint

A faint, stylized diagram in the background, consisting of a central vertical blue line with several horizontal and diagonal branches. The branches end in circles of various colors: blue, green, yellow, and orange. The diagram is centered behind the title text.

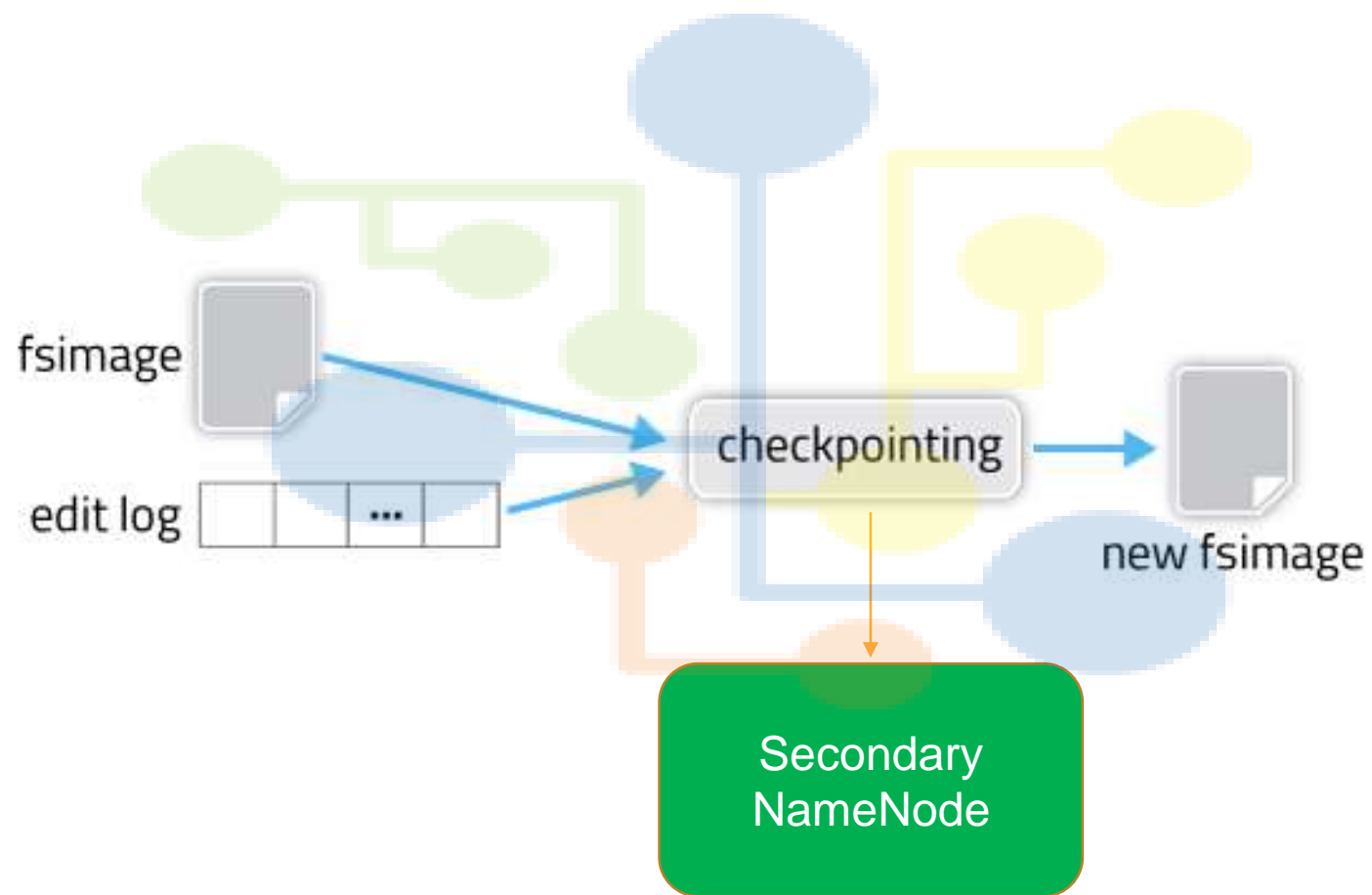


Checkpoint

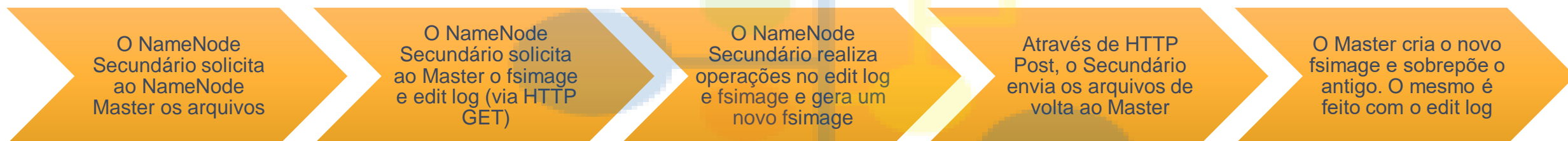


Data Science
Academy

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



Checkpoint é o processo pelo qual fsimage e edit log são combinados em um novo fsimage



O processo de checkpoint é controlado por 2 parâmetros

dfs.namenode.checkpoint.period	dfs.namenode.checkpoint.txns
O valor default é 1 hora, para garantir o próximo refresh entre 2 checkpoints consecutivos.	O valor default é 1 milhão e esta é a regra para definir o número de transações até o próximo checkpoint.

A faint, stylized diagram in the background of the slide, consisting of several colored circles (blue, green, yellow, orange) connected by lines, suggesting a network or data flow.

Compreender como funciona o checkpoint no HDFS pode fazer a diferença para ter um cluster eficiente.

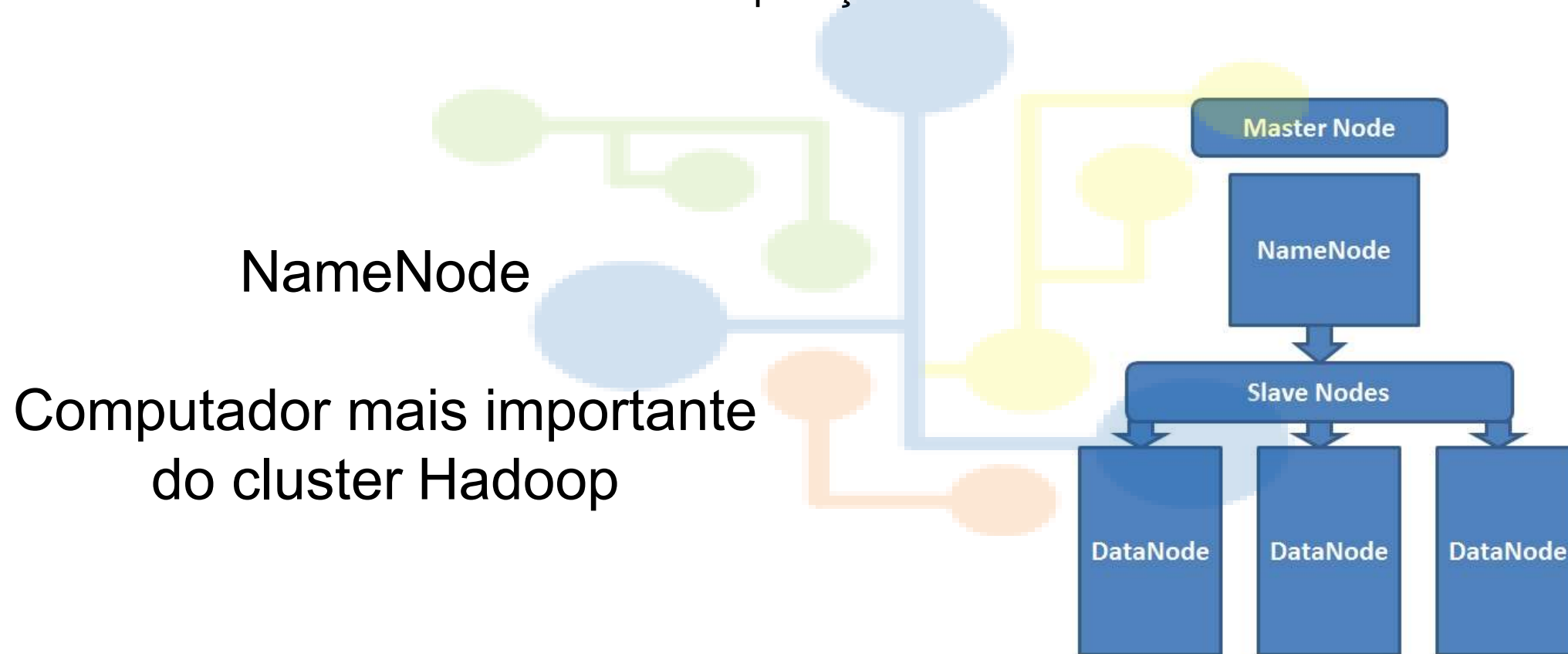


Procedimento de Recuperação à Falha do NameNode



Recuperação a Falhas

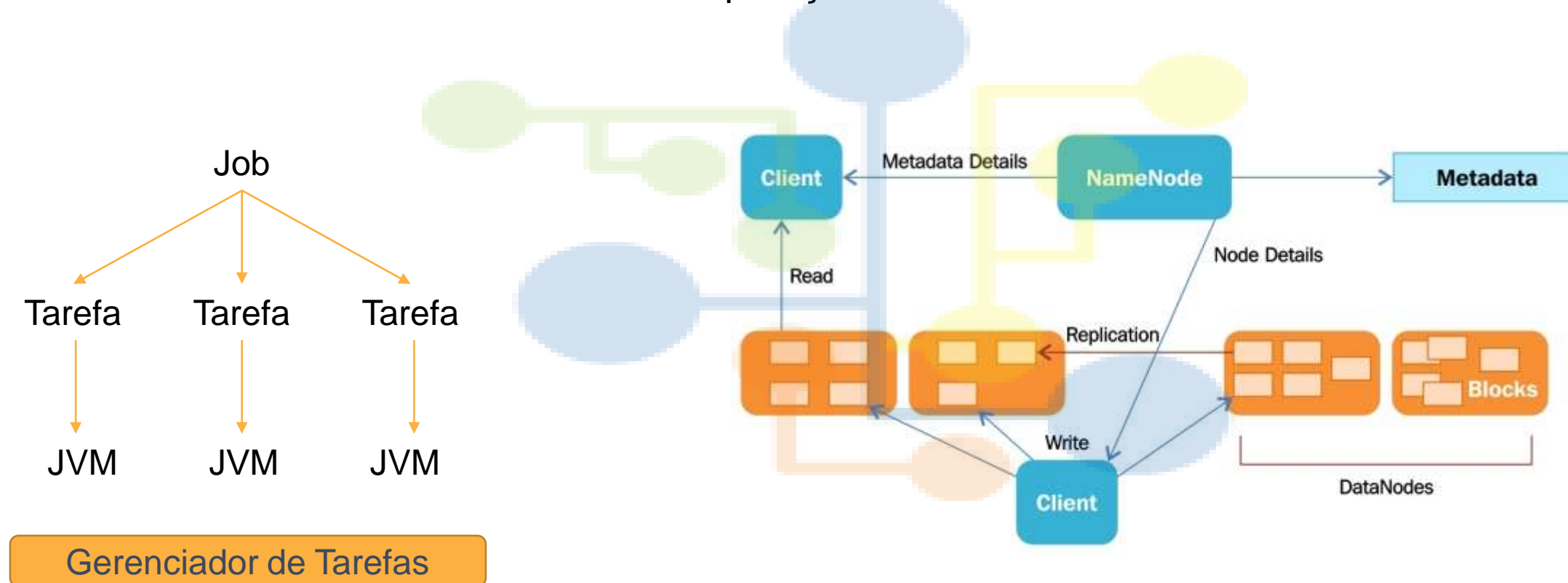
Processo de Recuperação à Falha do NameNode





Recuperação a Falhas

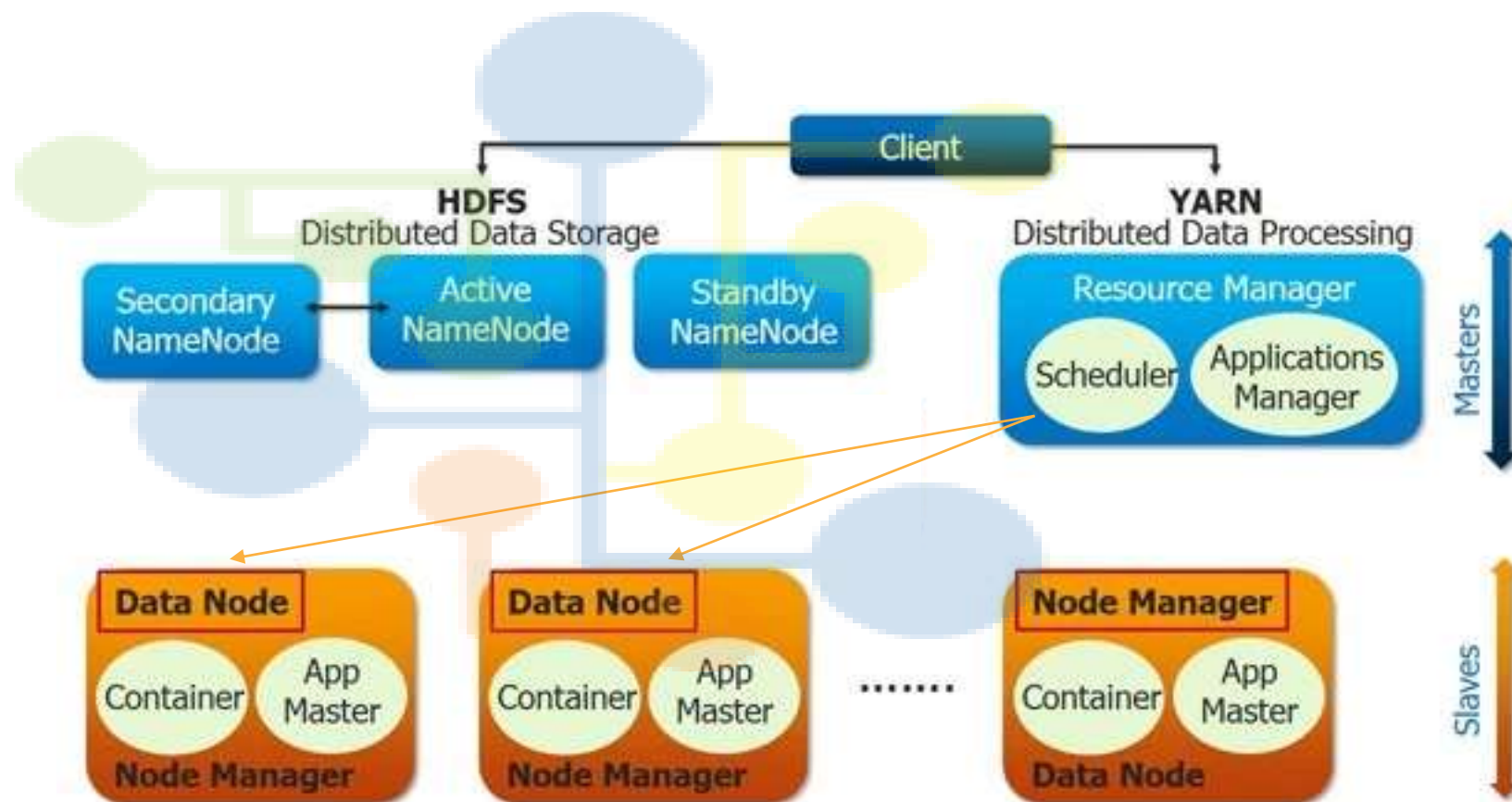
Processo de Recuperação à Falha do NameNode





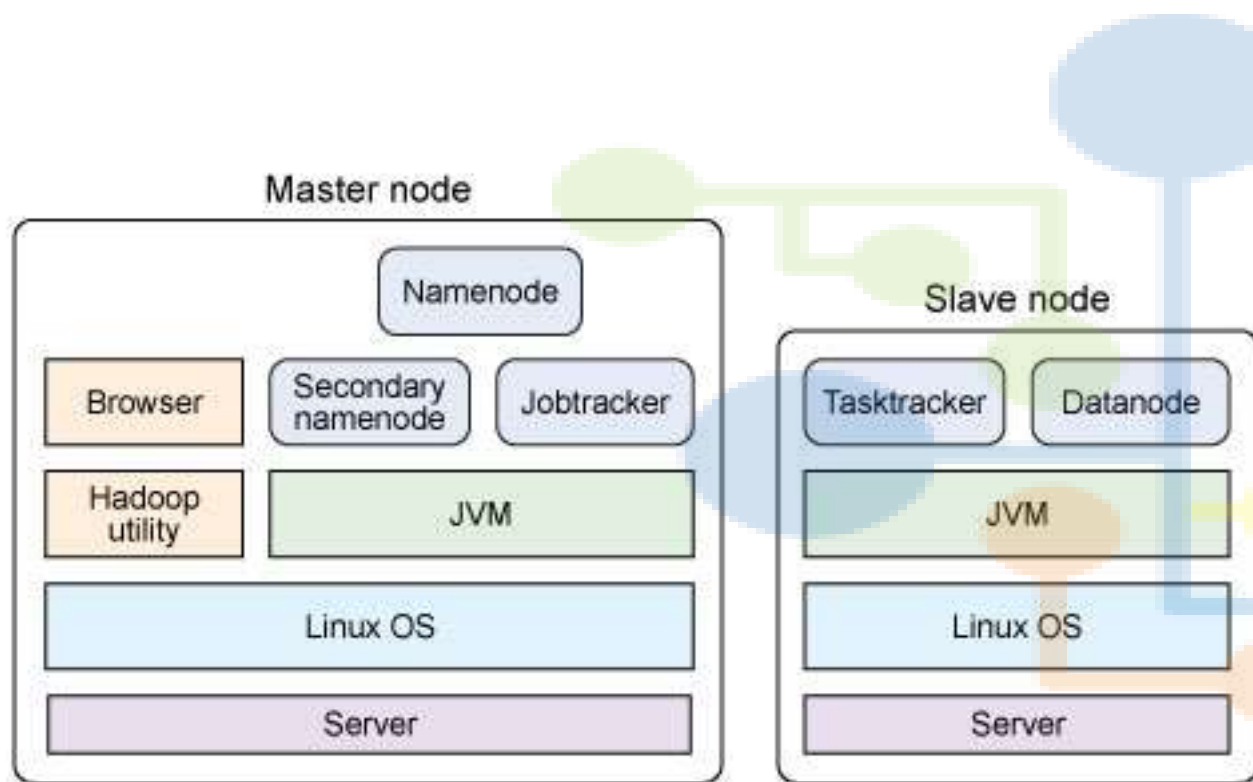
Recuperação a Falhas

Processamento de
Tarefas de Um Job
MapReduce





Recuperação a Falhas



O sucesso e segurança do processo de análise de Big Data com Hadoop, depende do bom funcionamento do Node Master



Recuperação a Falhas

Processo de Recuperação à Falha do NameNode

No caso de não haver backup do NameNode e o servidor falhar, o risco de perda de dados pode ser reduzido, criando um nível de redundância do NameNode

Faça uma cópia dos dados antes de promover o servidor a NameNode

Mude o endereço ip do novo servidor, para o endereço ip do servidor antigo

Garanta que o Hadoop esteja instalado e configurado de forma idêntica ao original

Não formate o NameNode



Modo de Segurança



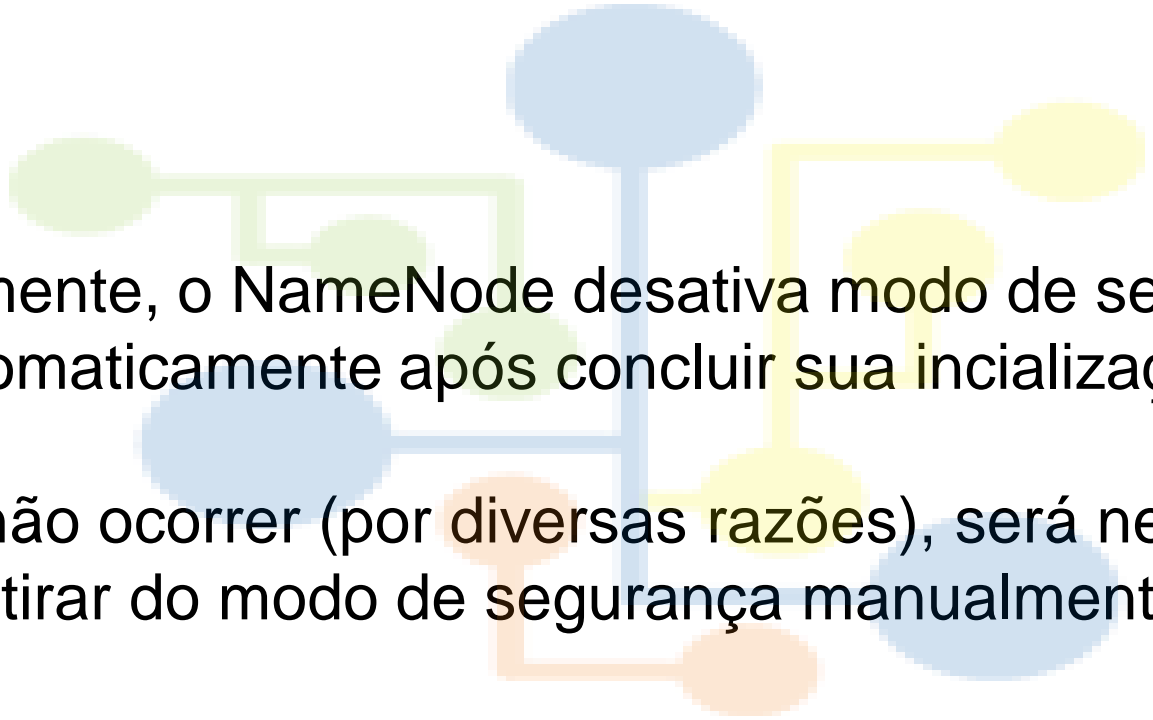


Modo de Segurança

Modo de Segurança (Safe Mode) é o modo apenas leitura do cluster HDFS, onde modificações não são permitidas no file system ou nos blocos.



Modo de Segurança

A faint, stylized diagram in the background consisting of several colored circles (blue, green, yellow, orange) connected by thin lines, resembling a network or flowchart.

Normalmente, o NameNode desativa modo de segurança automaticamente após concluir sua inicialização.

Se isso não ocorrer (por diversas razões), será necessário retirar do modo de segurança manualmente.



Modo de Segurança

Safe Mode

Em Safe Mode:

- Somente operações no file system, que acessam os metadados podem ser executadas
- Leitura de arquivos serão possíveis apenas se os blocos estiverem disponíveis nos DataNodes
- Modificações em arquivos não serão efetivadas

Para colocar o HDFS em Safe Mode, use o comando:

```
hdfs dfsadmin -safemode
```

```
hdfs dfsadmin -safemode leave
```




Backup

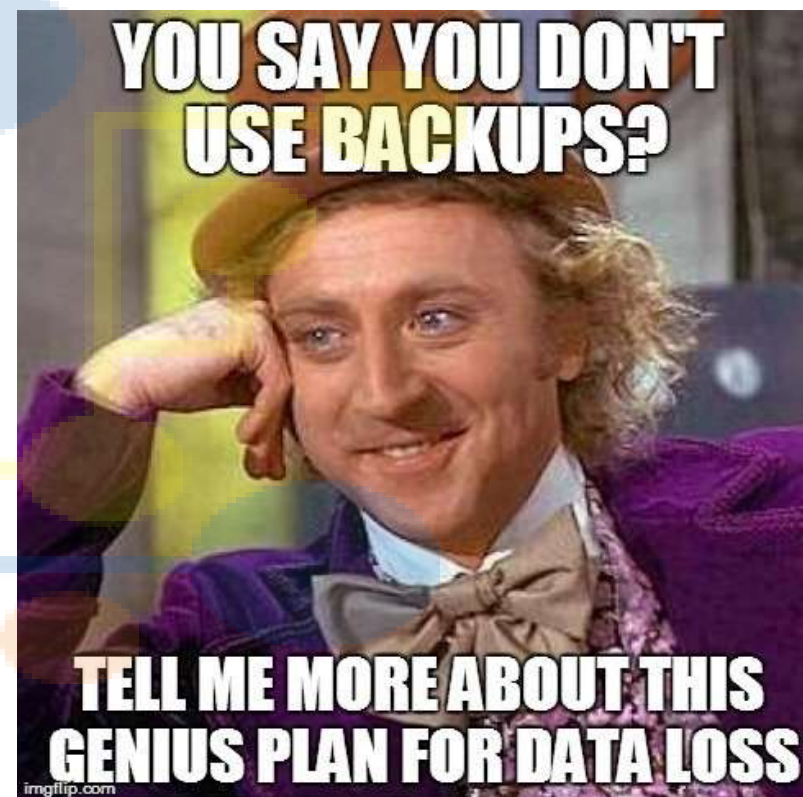
Backup



Data Science
Academy

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

Backup



Backup



Data Science
Academy

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

No Hadoop o backup é composto de 2 partes:

- Backup dos metadados
- Backup dos dados

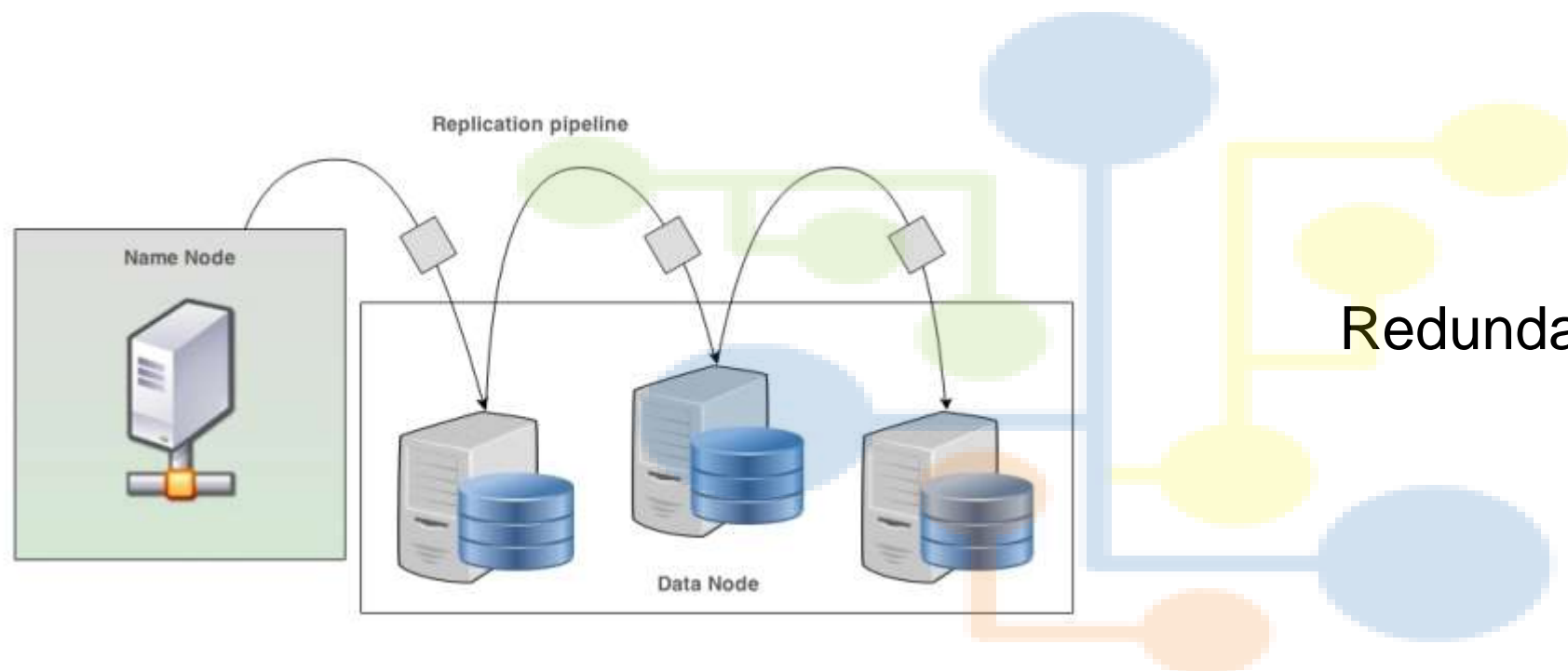


Backup



Data Science
Academy

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



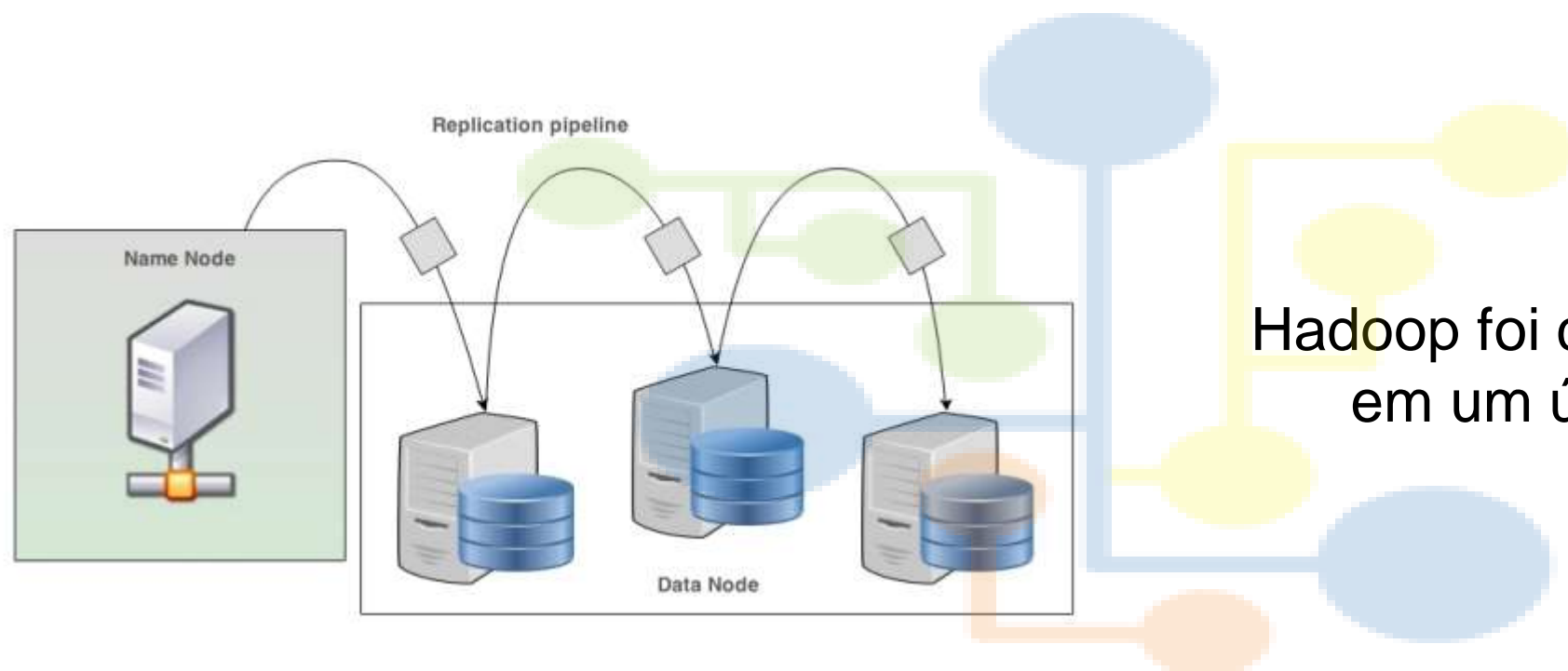
Redundante a Falhas por
Padrão

Backup



Data Science
Academy

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



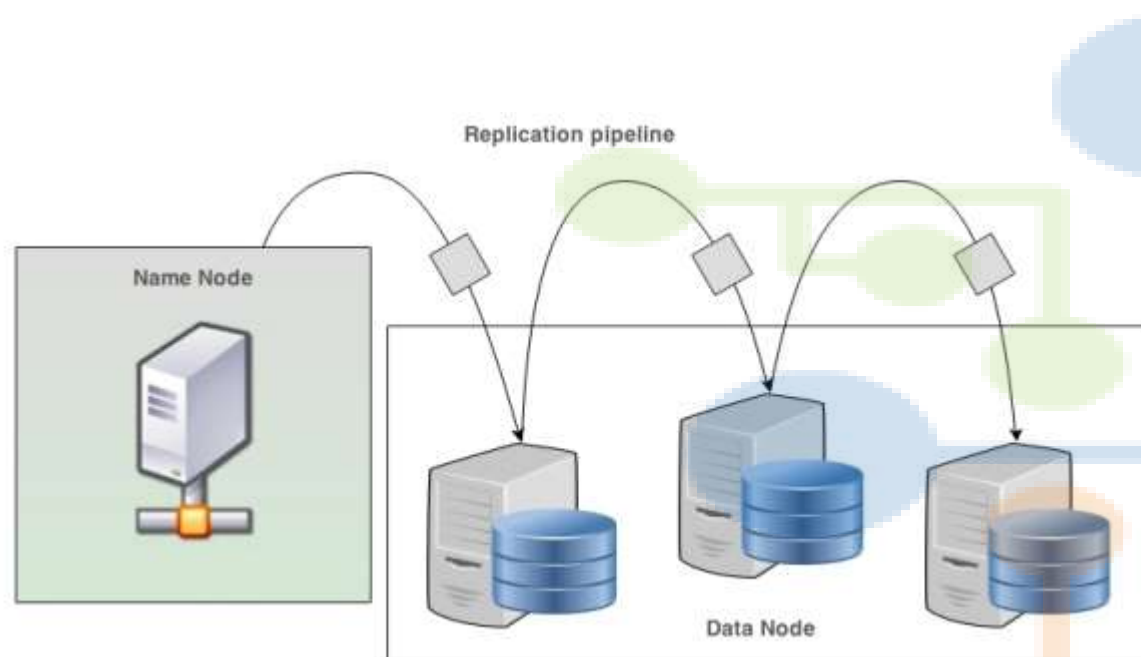
Hadoop foi criado para executar
em um único Datacenter

Backup



Data Science
Academy

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



Hadoop permite o
armazenamento de dados
compactados

Parquet e ORC Files

Backup do HDFS

Soluções
Proprietárias

Replicação do
HDFS

Cópia dos
Dados

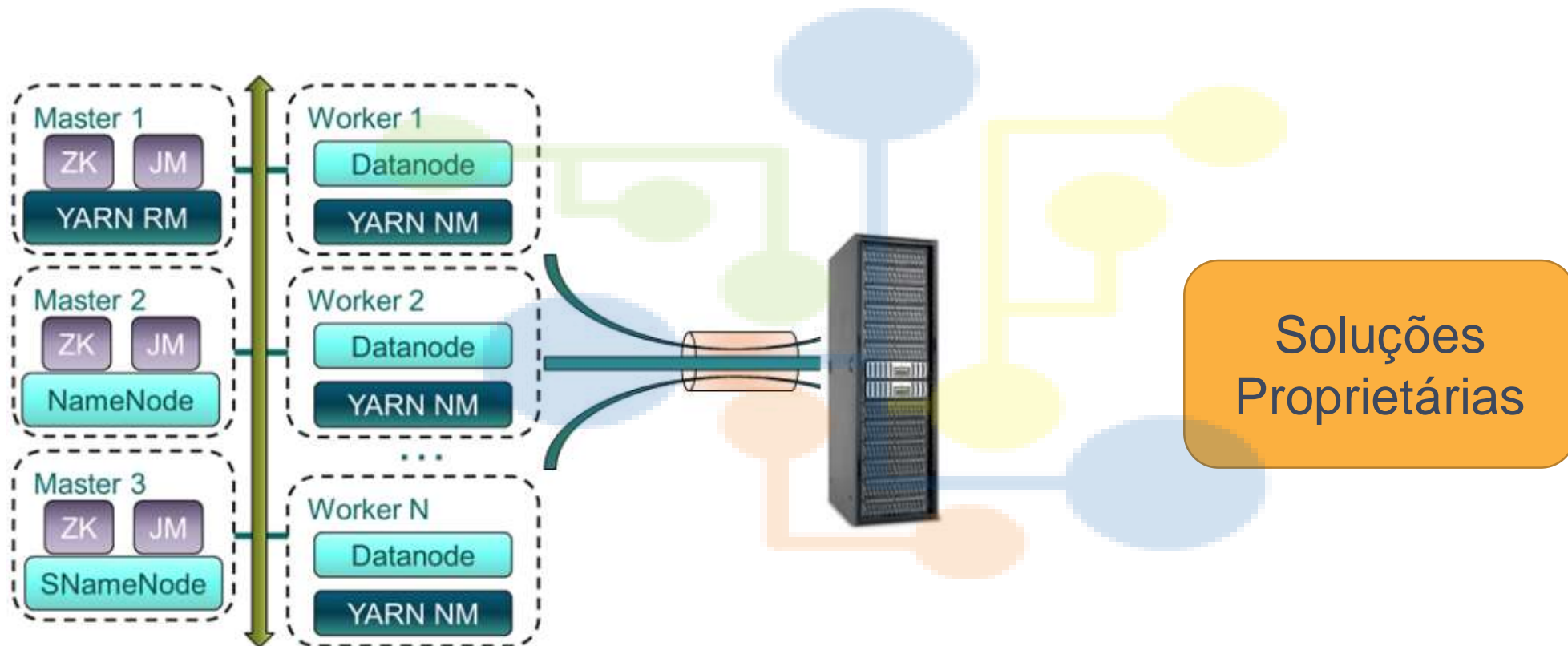
Dual Load

Backup



Data Science
Academy

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



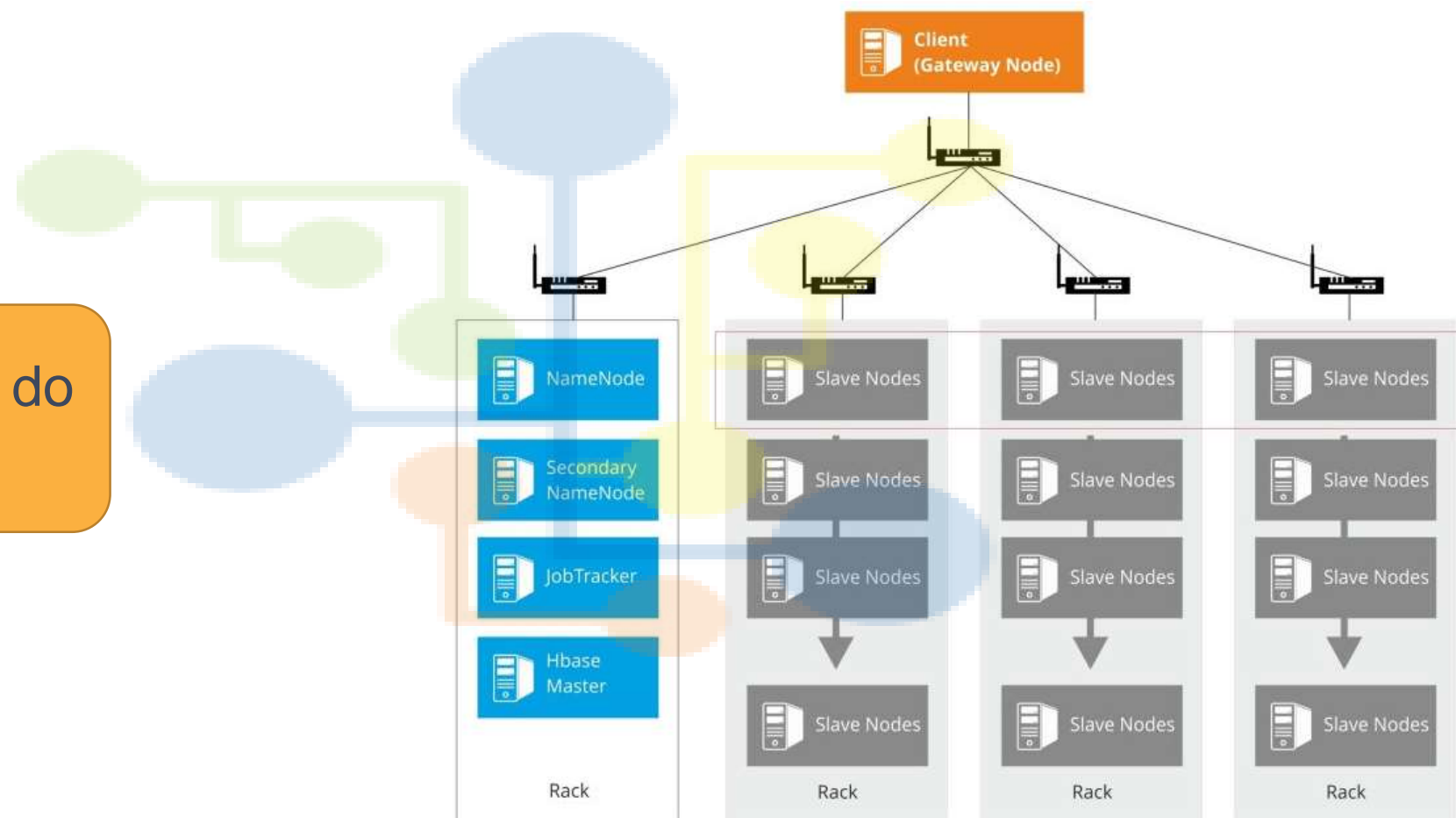
Backup



Data Science
Academy

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

Replicação do HDFS

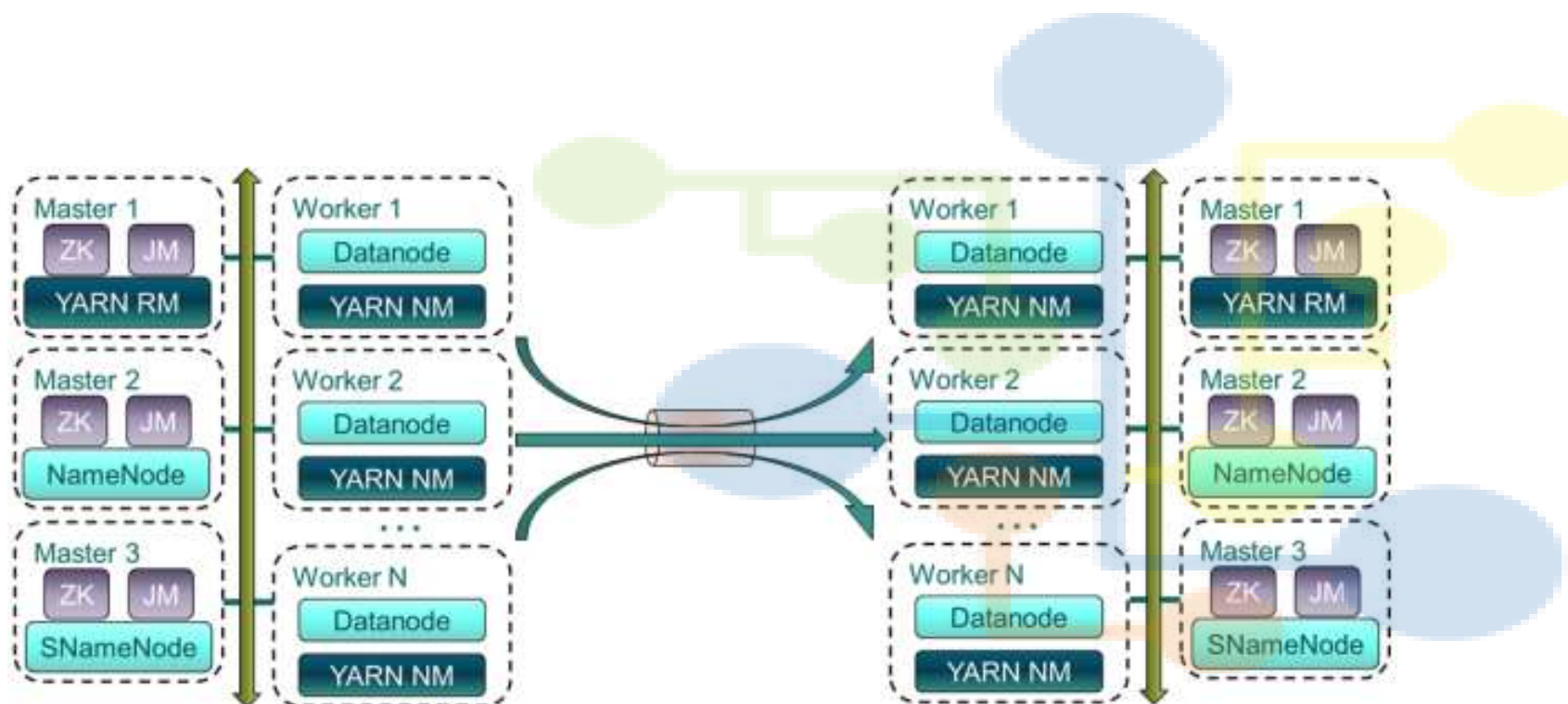


Backup



Data Science
Academy

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



Cópia dos
Dados

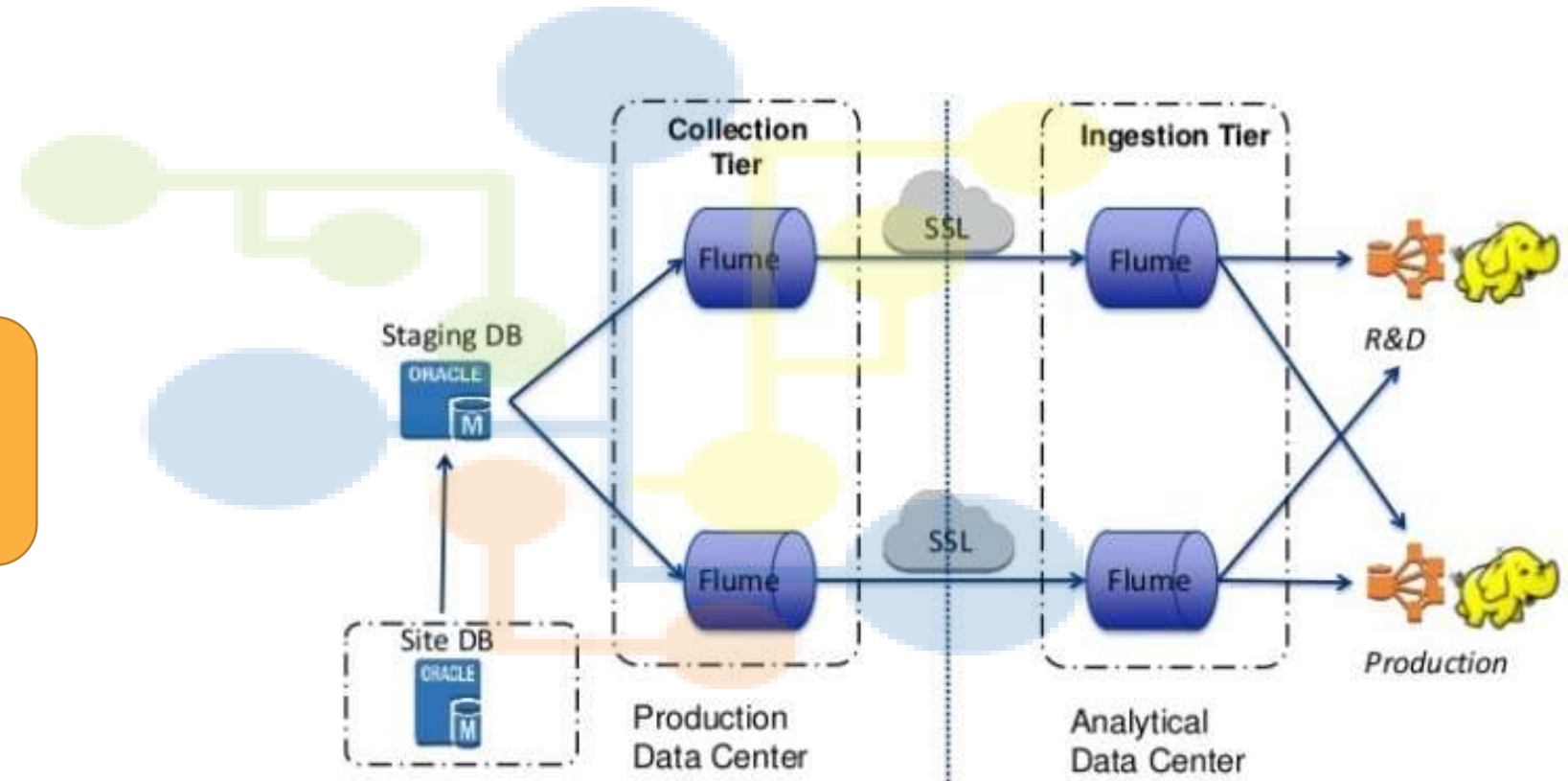
Backup



Data Science
Academy

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

Dual Load

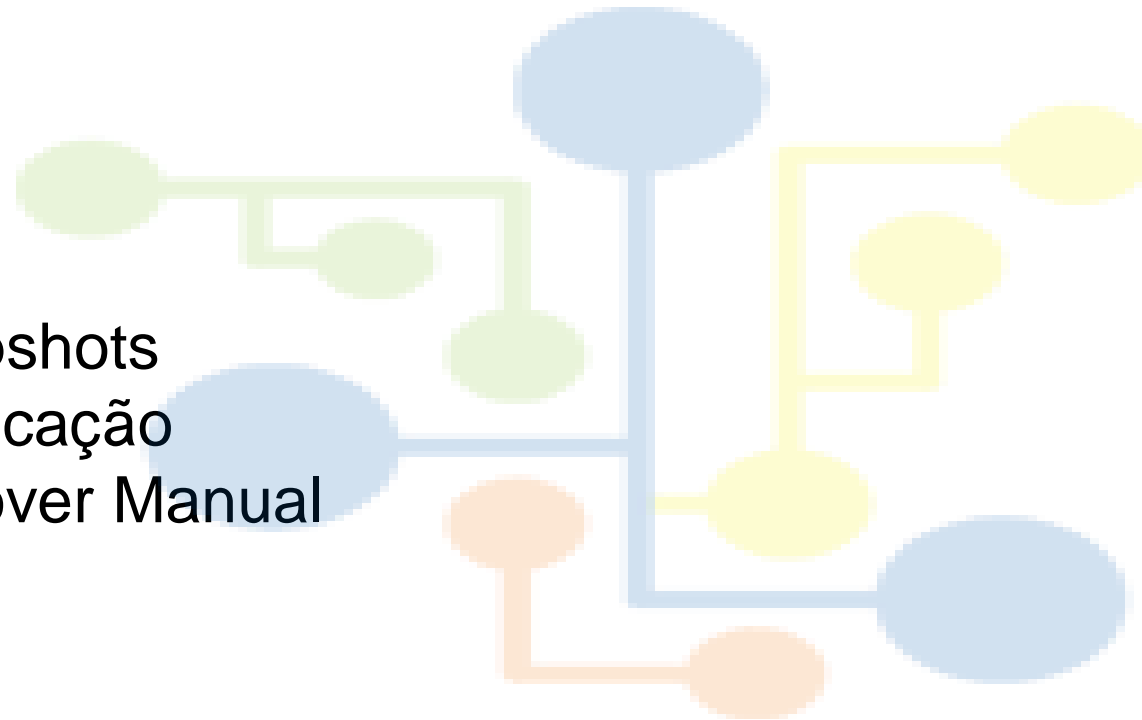


Apenas o Backup é suficiente?

- Divisão dos dados em críticos e não críticos
- Formato do storage que será usado com o HDFS
- Monitoramento do Cluster
- Aplicação de Patches e Correções de Segurança

Recovery

- Snapshots
- Replicação
- Recover Manual
- API



Backup



Data Science
Academy

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

Filesystem check

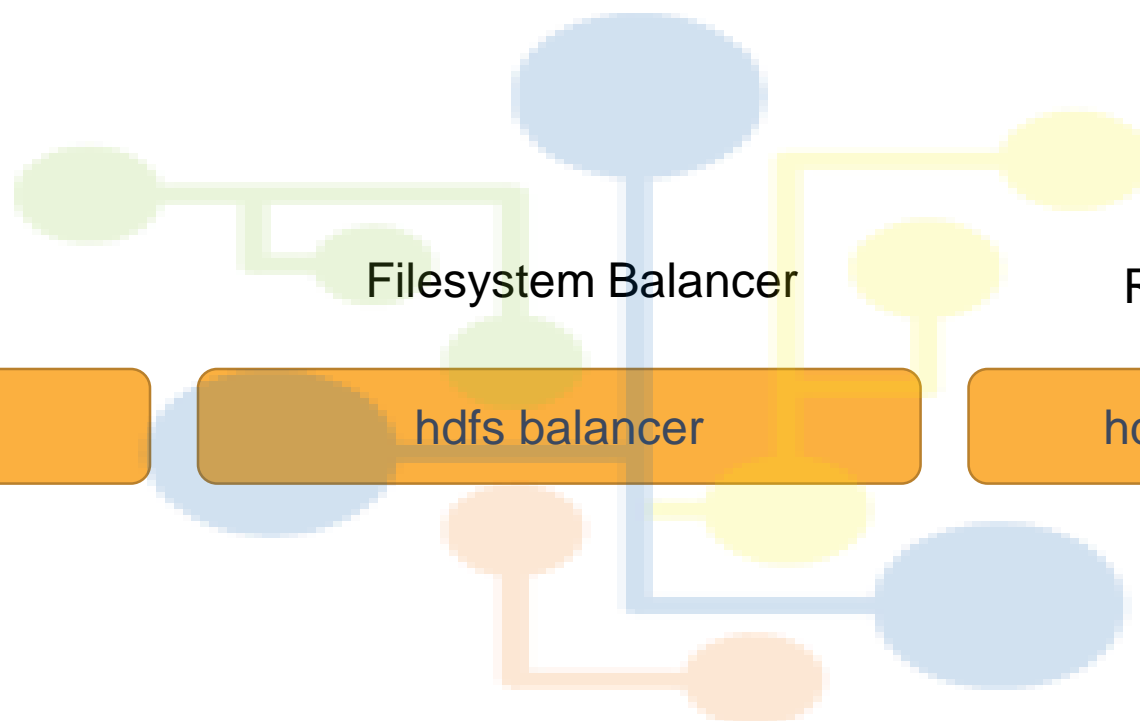
`hdfs fsck /`

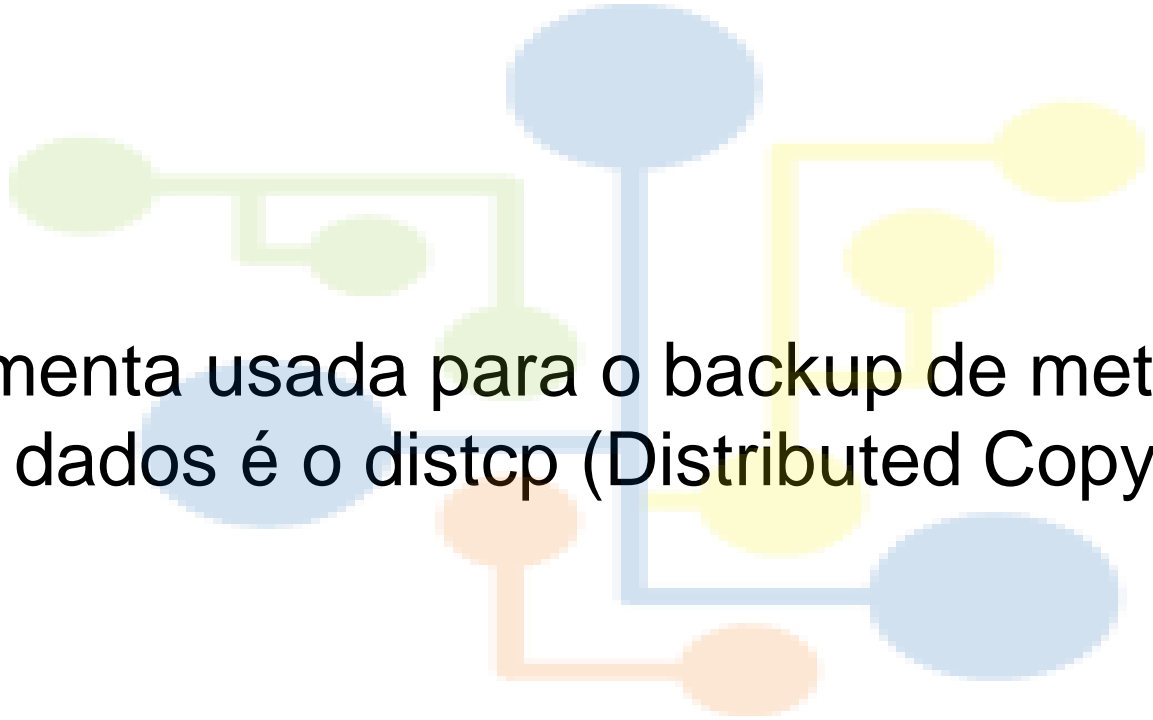
Filesystem Balancer

`hdfs balancer`

Relatório do cluster

`hdfs dfsadmin -report`



A faint, stylized diagram in the background consisting of several colored circles (blue, green, yellow, orange) connected by lines, representing a network or data flow.

A ferramenta usada para o backup de metadados e dados é o distcp (Distributed Copy).

Backup



Data Science
Academy

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

Metadados dos demais produtos do ecossistema Hadoop, também devem ser incluídos no procedimento de Backup



Solução de Problemas no Cluster Hadoop





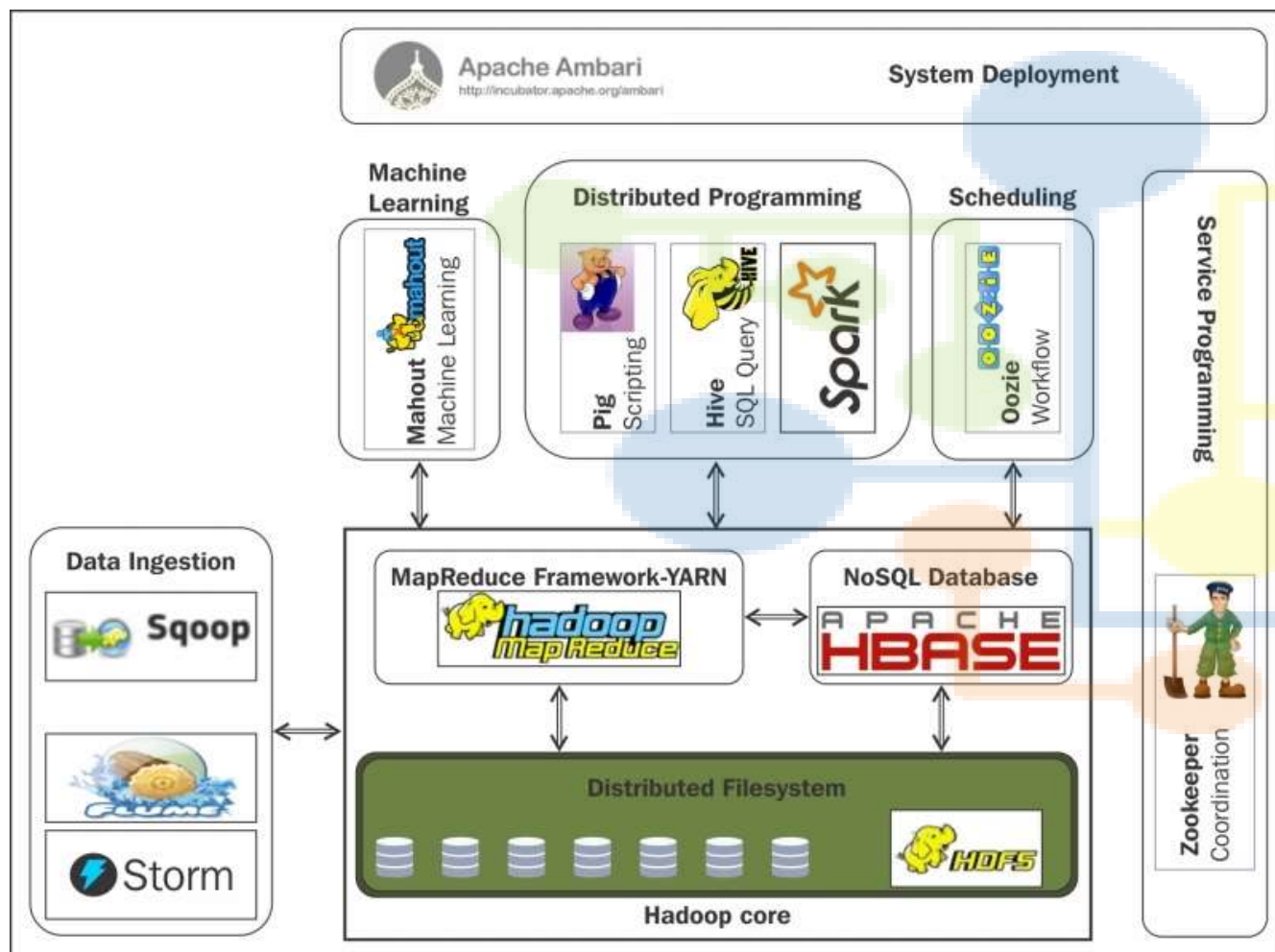
Solução de Problemas no Hadoop

O Hadoop não é um banco de dados e na condição de repositório de dados, ele possui muitas vantagens em termos de gerenciamento e segurança





Solução de Problemas no Hadoop



Hadoop é um grande ecossistema



Solução de Problemas no Hadoop

Possíveis problemas na Administração do Hadoop

Erro humano

Erros humanos são uma das principais causas de problemas com o Hadoop

Hardware

Falhas de disco ocorrem com frequência e nem sempre de uma vez. A degradação ao longo do tempo, pode levar à lentidão, antes de uma falha ocorrer

Erros de configuração

É difícil configurar um ambiente Hadoop 100% otimizado e são muitos parâmetros possíveis. A utilização destes parâmetros depende de uma série de fatores por isso recomenda-se a realização maciça de testes antes de aplicar alterações em produção



Solução de Problemas no Hadoop

Possíveis problemas na Administração do Hadoop

Excessiva Utilização de Recursos

Erros em tarefas devem ser investigados e resolvidos. Erros recorrentes, consomem recursos do servidor, degradando performance

Identificação dos nodes

Configuração de rede dos nodes é um item crítico, pois a comunicação entre eles poderá gerar problemas de performance



Solução de Problemas no Hadoop

Solução de Problemas em um Ambiente Hadoop





Autenticação e Segurança no Hadoop



Autenticação e Segurança no Hadoop

Riscos de Segurança no Armazenamento de Big Data

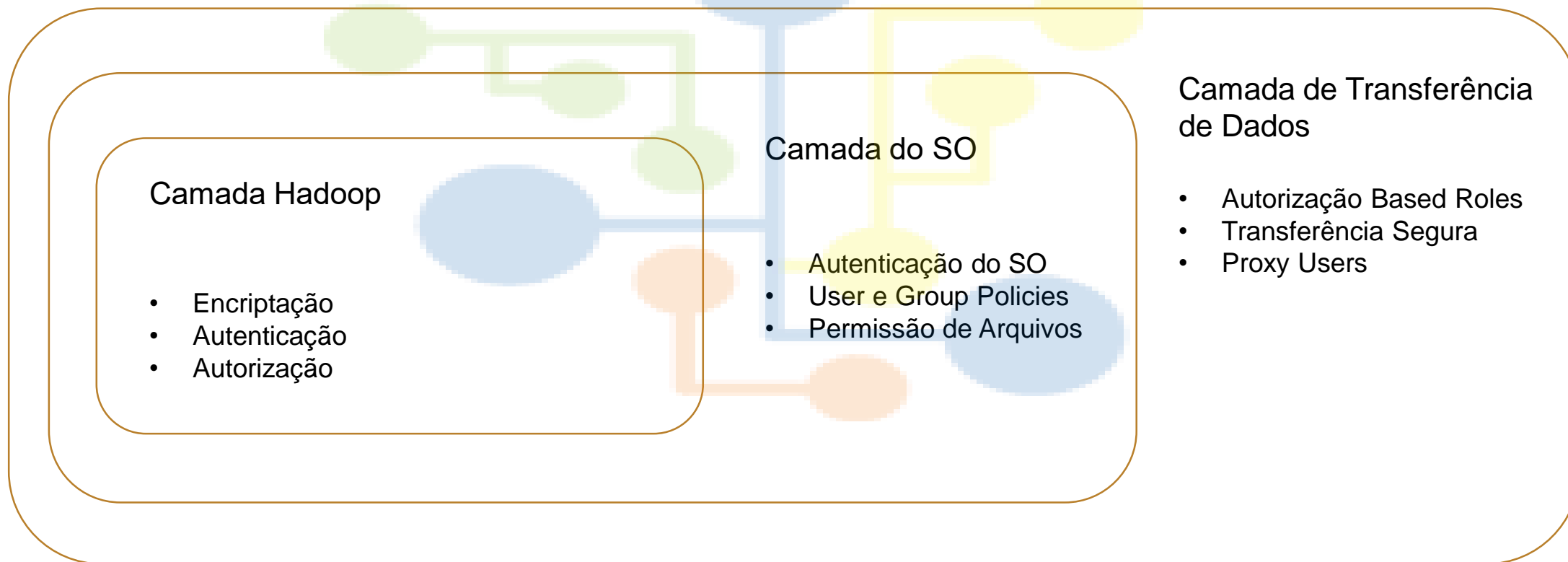
- Perda de Dados
- Queda de Produtividade
- Redução do Valor Geral dos Dados
- Vazamento de Dados Confidenciais





Autenticação e Segurança no Hadoop

A segurança no Hadoop é feita através de múltiplas camadas





Autenticação e Segurança no Hadoop

O Hadoop possui 2 métodos de autenticação





Autenticação e Segurança no Hadoop

Protocolo Kerberos



© Allison Smith, Amosink
Interactive week



Autenticação e Segurança no Hadoop

Autenticação via Kerberos

Como funciona o Kerberos

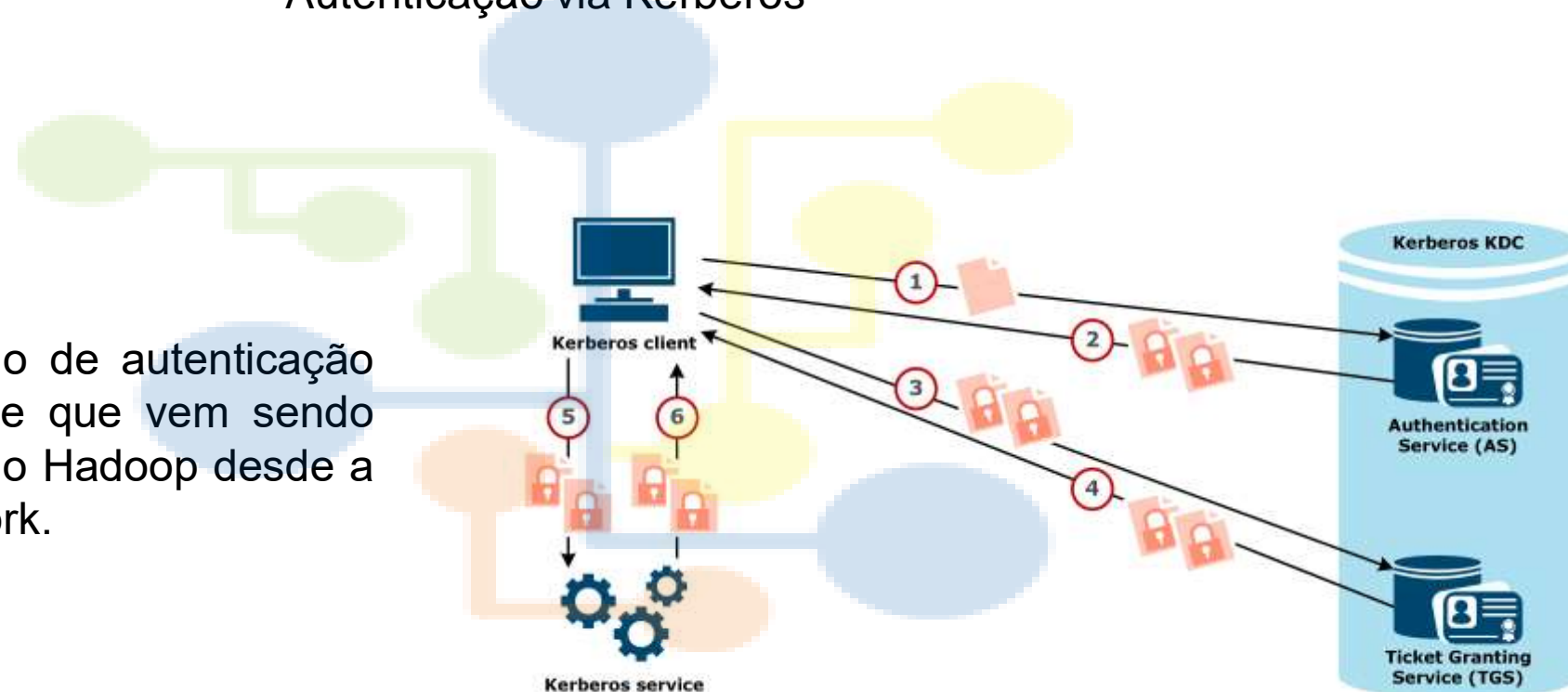




Autenticação e Segurança no Hadoop

Autenticação via Kerberos

Kerberos é um protocolo de autenticação de redes, open-source e que vem sendo usado na autenticação do Hadoop desde a versão 0.20 do Framework.

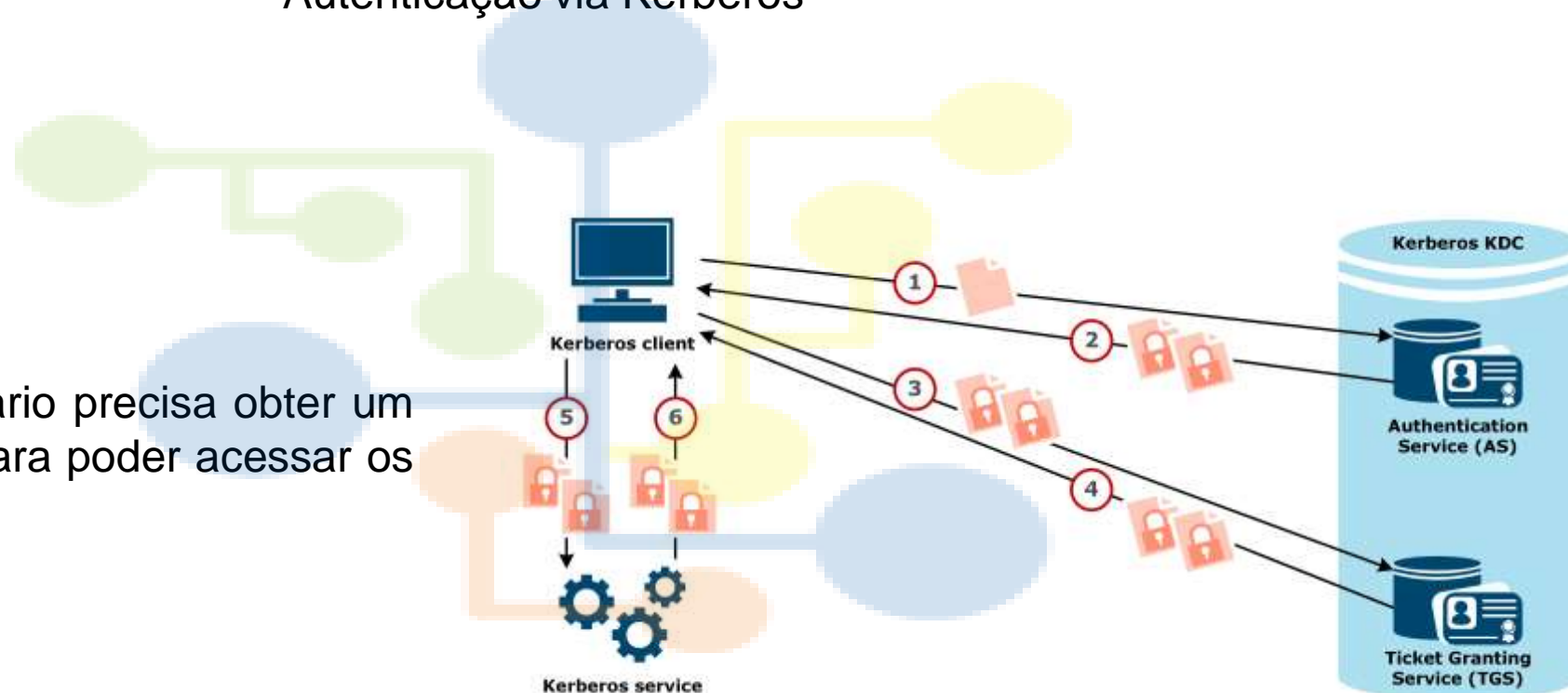




Autenticação e Segurança no Hadoop

Autenticação via Kerberos

Com o Kerberos, o usuário precisa obter um token de autenticação para poder acessar os dados no HDFS.





Autenticação e Segurança no Hadoop

Configuração do Kerberos

1

- Instalar os pacotes do Kerberos no sistema operacional
- `yum install krb5-server krb5-libs krb5-auth-dialog krb5-workstation`

2

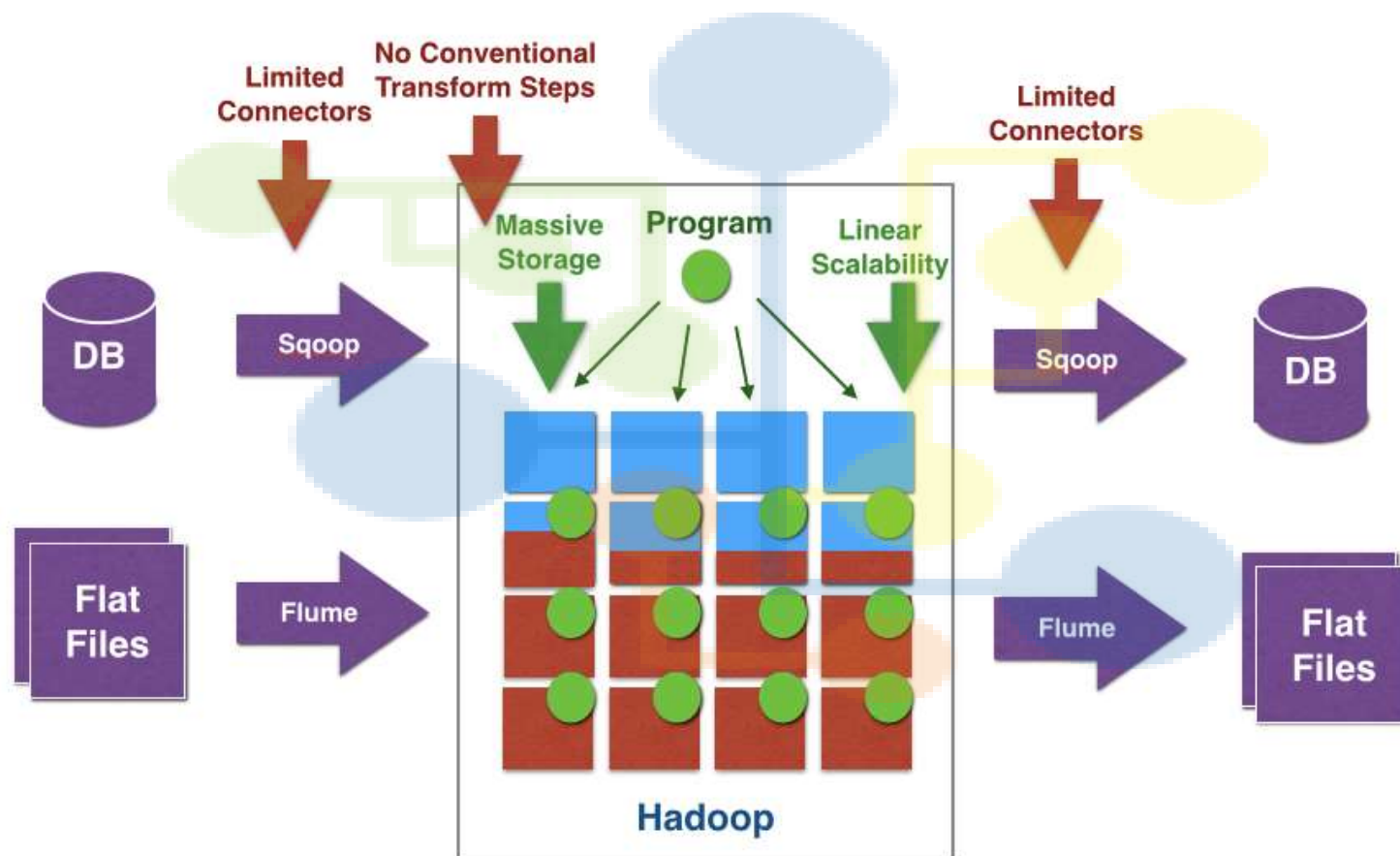
- Editar os arquivos de configuração
- `/etc/krb5.conf`
- `/var/kerberos/krb5kdc/kdc.conf`

3

- Copiar as versões atualizadas dos arquivos para cada node no cluster



Autenticação e Segurança no Hadoop



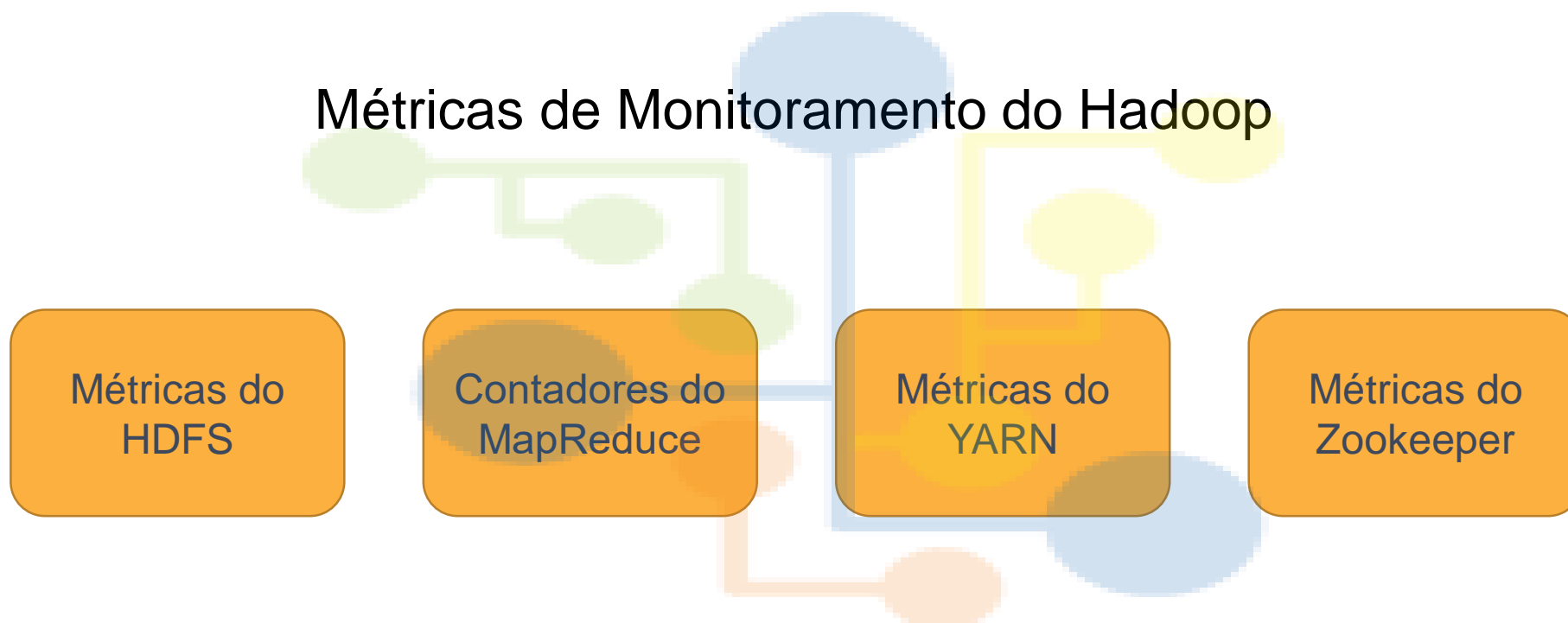


Melhores Práticas de Monitoramento do Cluster Hadoop



Monitoramento

Métricas de Monitoramento do Hadoop



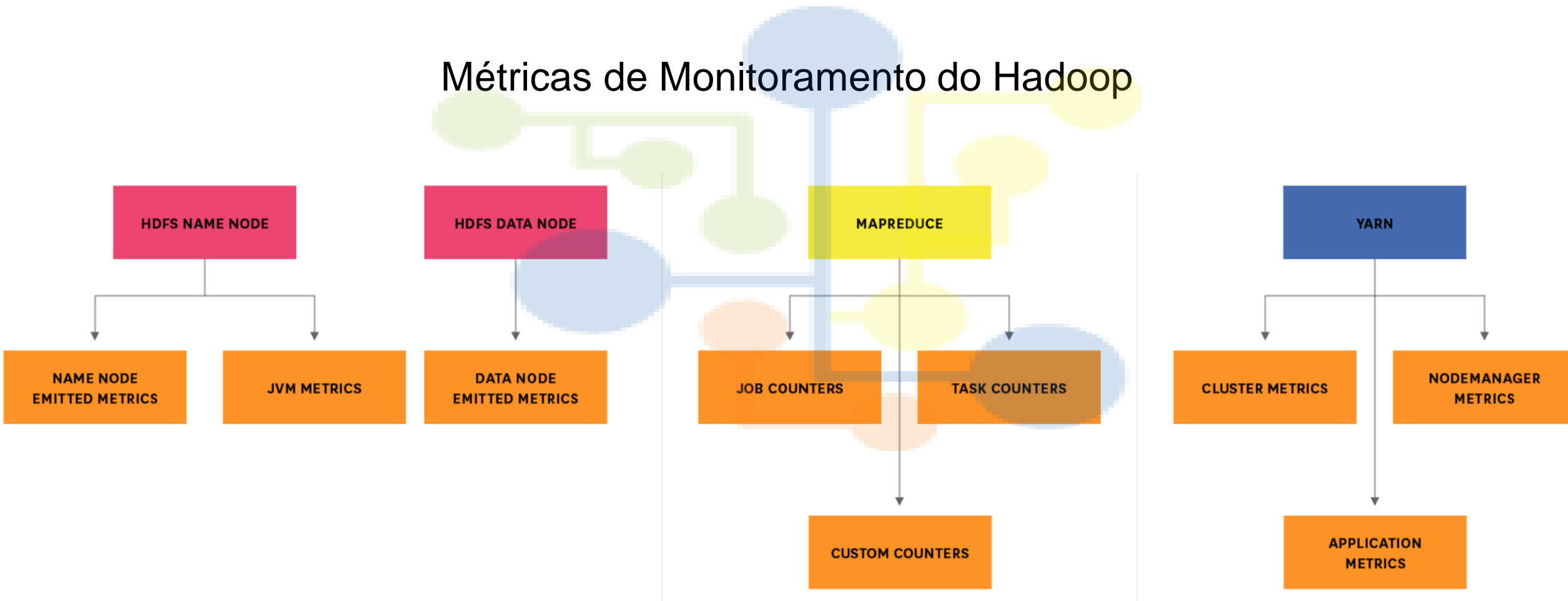
Monitoramento



Data Science
Academy

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

Métricas de Monitoramento do Hadoop



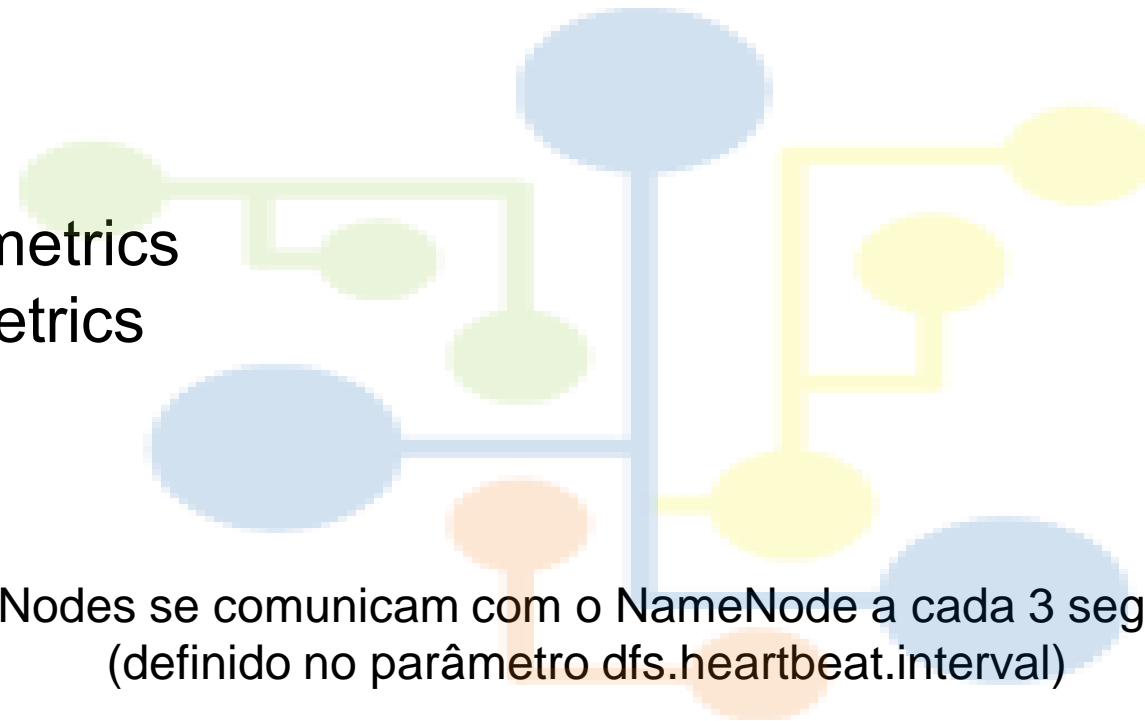


Monitoramento

Métricas HDFS

- NameNode metrics
- DataNode metrics

DataNodes se comunicam com o NameNode a cada 3 segundos
(definido no parâmetro `dfs.heartbeat.interval`)





Monitoramento

Métricas HDFS

- NameNode metrics
- DataNode metrics

Métrica a ser coletada
CapacityRemaining
CorruptBlocks / MissingBlocks
VolumeFailuresTotal
NumLiveDataNodes / NumDeadDataNodes
FilesTotal
TotalLoad
BlockCapacity / BlocksTotal
UnderReplicatedBlocks
NumStaleDataNodes



Contadores MapReduce

- Job counters
- Task counters
- Custom counters
- File system counters

Métrica a ser coletada

MILLIS_MAPS/MILLIS_REDUCES

NUM_FAILED_MAPS/NUM_FAILED_REDUCES

REDUCE_INPUT_RECORDS

SPILED_RECORDS

GC_TIME_MILLIS

NUM_FAILED_MAPS

NUM_FAILED_REDUCES

RACK_LOCAL_MAPS

DATA_LOCAL_MAPS



Monitoramento

Métricas YARN

- Cluster metrics
- Application metrics
- NodeManager metrics

Métrica a ser coletada

unhealthyNodes

activeNodes

lostNodes

appsFailed

totalMB / allocatedMB

progress

containersFailed



Monitoramento

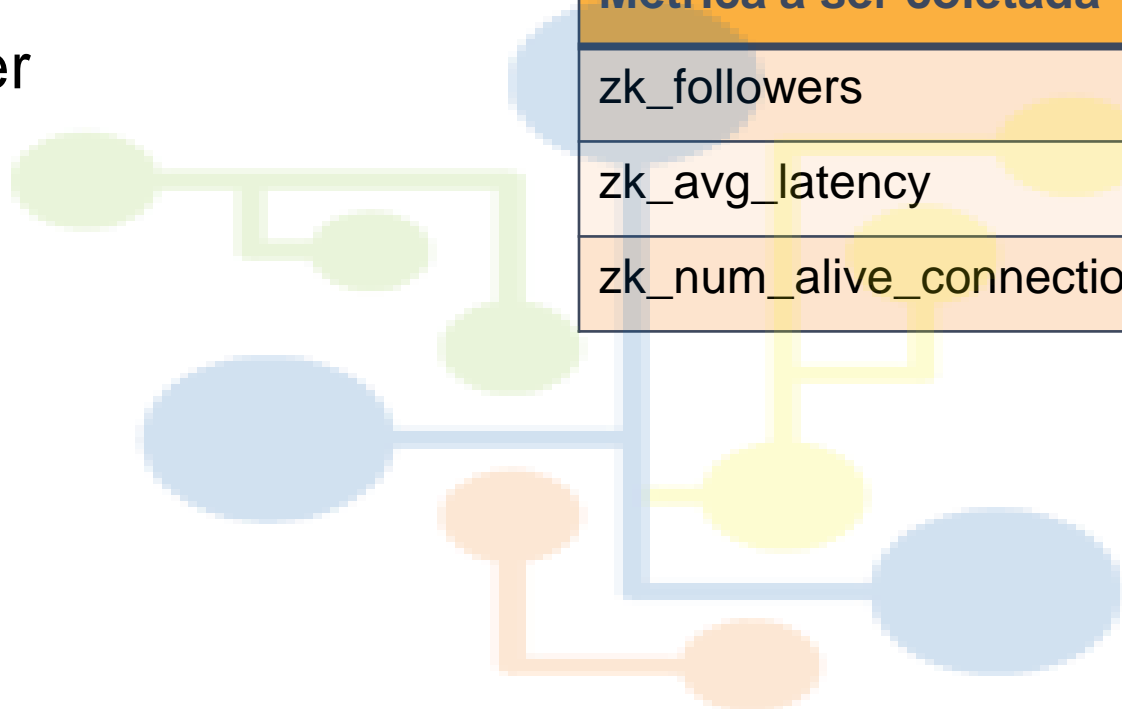
Métricas Zookeeper

Métrica a ser coletada

zk_followers

zk_avg_latency

zk_num_alive_connections





Monitoramento

Como coletar as métricas?

- NameNode → coleta de métricas via API
- DataNode → coleta de métricas via API
- HDFS → coleta de métricas via JMX



Monitoramento

Como coletar as métricas?

- NameNode → coleta de métricas via API

`http://<namenodehost>:50070`



Como coletar as métricas?

- DataNode → coleta de métricas via API

<http://datanodehost:50070/dfshealth.html#tab-datanode>



Monitoramento

Como coletar as métricas?

- HDFS → coleta de métricas via jmx

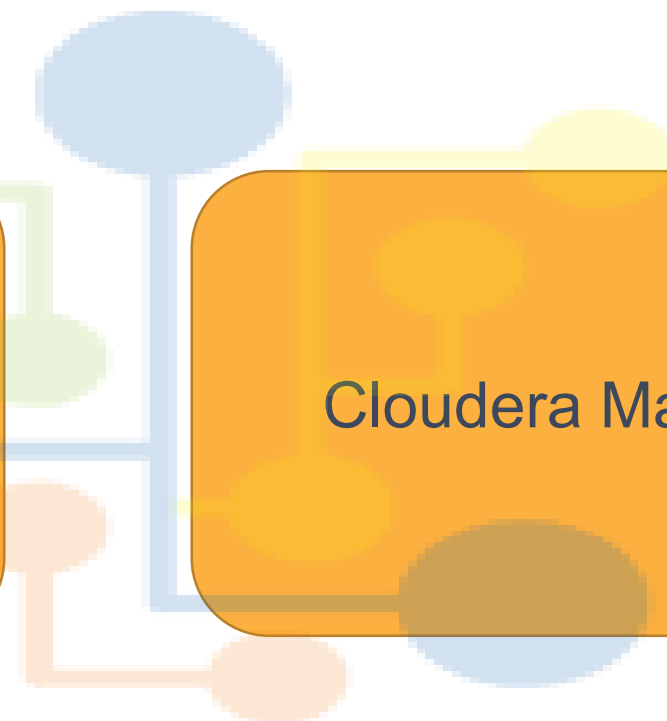
`http://<namenodehost>:50070/jmx`



Monitoramento

Apache Ambari

Cloudera Manager





Monitoramento

Apache Ambari

Utilizado pelo:

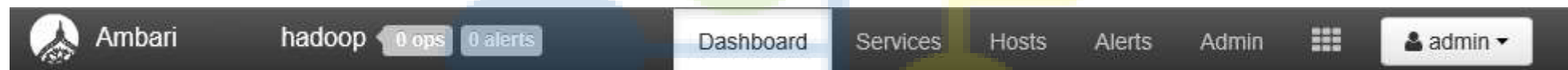
- Microsoft HDInsight
- Hortonworks

Monitoramento



Data Science
Academy

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



Monitoramento



Data Science
Academy

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

Dashboard



Monitoramento



Data Science
Academy

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

Grupos de Alerta

Manage Alert Groups

You can manage alert groups for each service in this dialog. View the list of alert groups and the alert definitions configured in them. You can also add/remove alert definitions, and pick notification for that alert group.

AMBARI Default (44)	Metric Collector HBase Maser CPU Utilization
AMS Default (44)	Metric Collector - HBase Master Process
HDFS Default (44)	Metric Collector Process
HIVE Default (44)	Metric Monitor Status
KAFKA Default (44)	Percent Metric Monitors Available
MAPREDUCE2 Default (44)	Metric Collector - ZooKeeper Server Process
OOZIE Default (44)	History Server Web UI
PIG Default (44)	History Server Process
SQOOP Default (44)	History Server RPC Latency
TEZ Default (44)	History Server CPU Utilization
YARN Default (44)	Secondary NameNode Process
ZOOKEEPER Default (44)	NameNode High Availability Health
	DataNode Web UI
	NameNode Host CPU Utilization
	NameNode RPC Latency

+

-

⚙

☒ Notifications

Add

Cancel

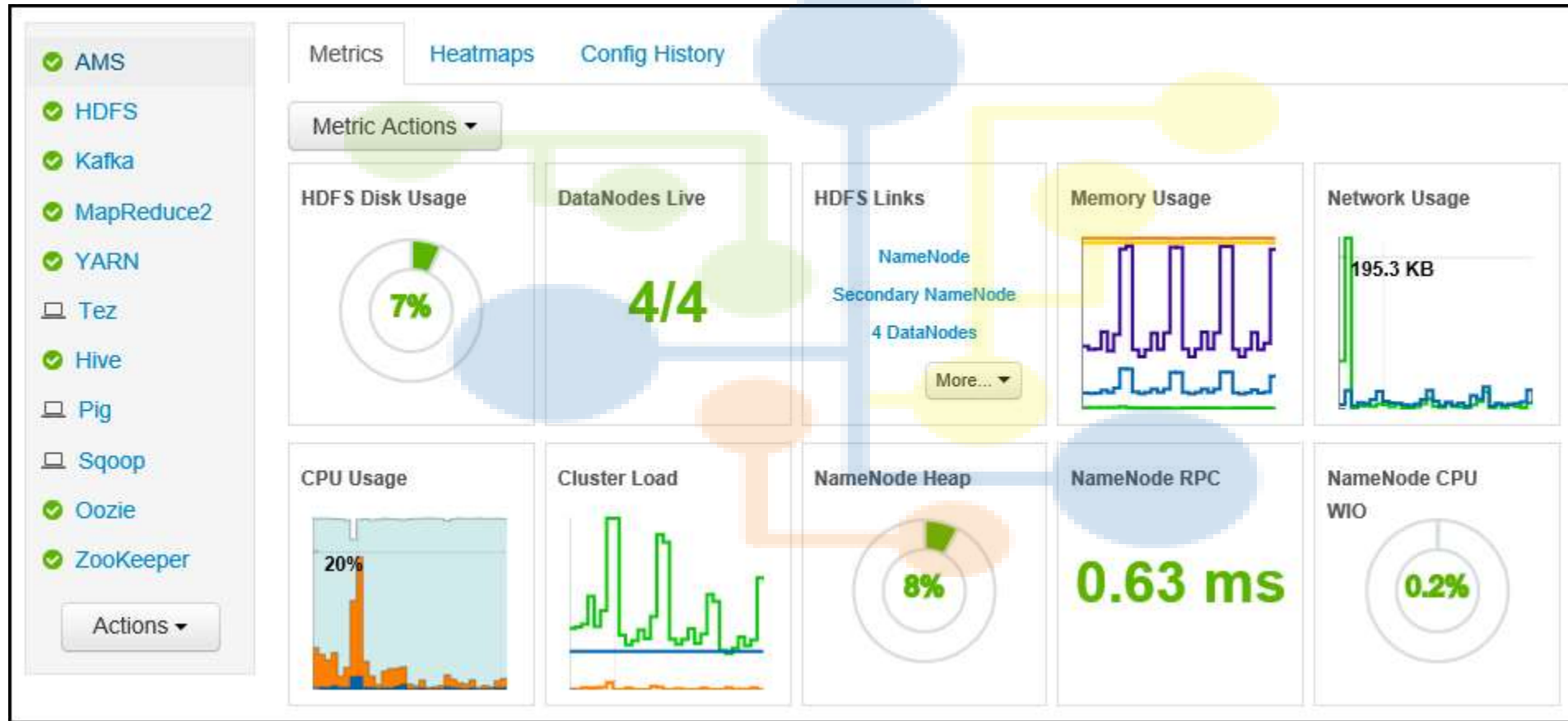
Save

Monitoramento



Data Science
Academy

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





Obrigado

