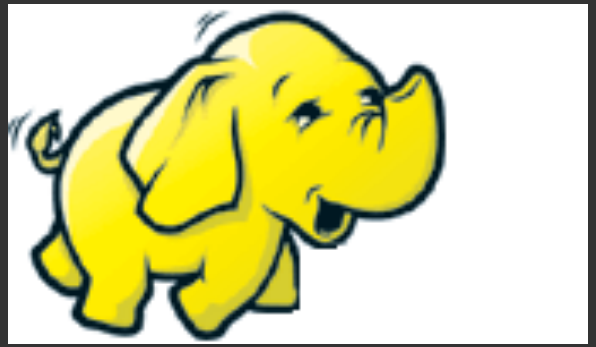
HDFS@toutiao

徐鹏 Data-Inf

01 Intro

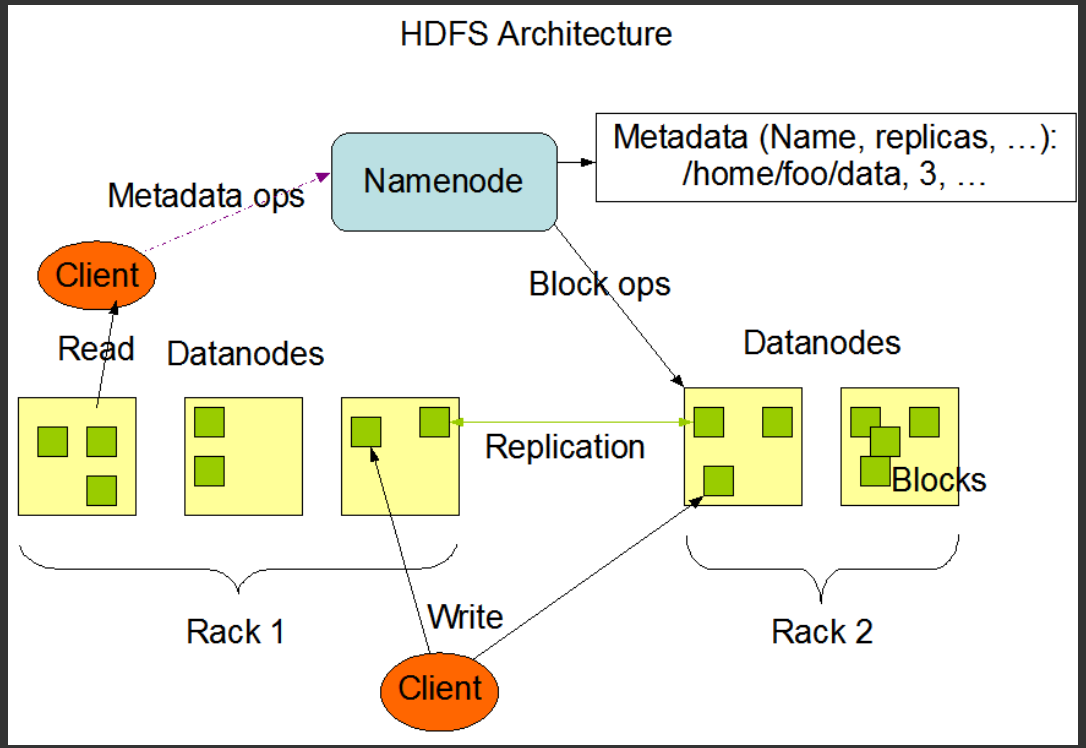
Introduction



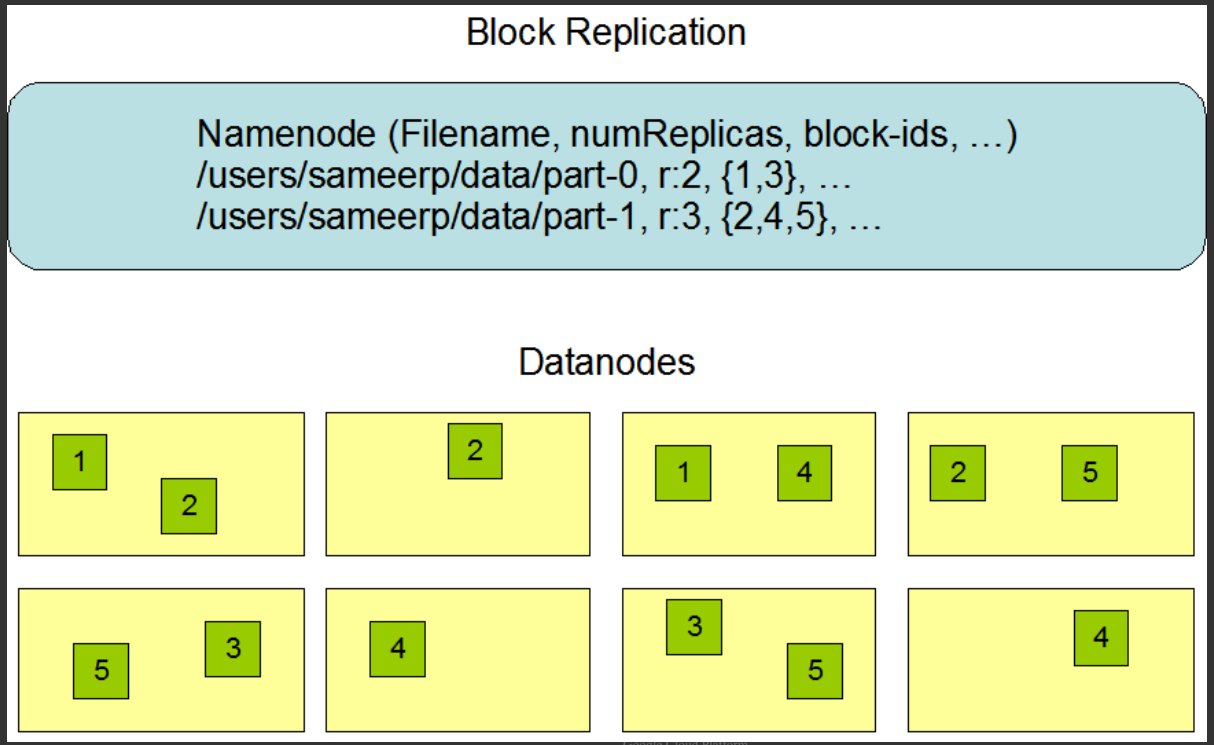
- Java, Apache

- Large Data Sets
- scalable, fault-tolerant, distributed storage system
- high throughput of data access

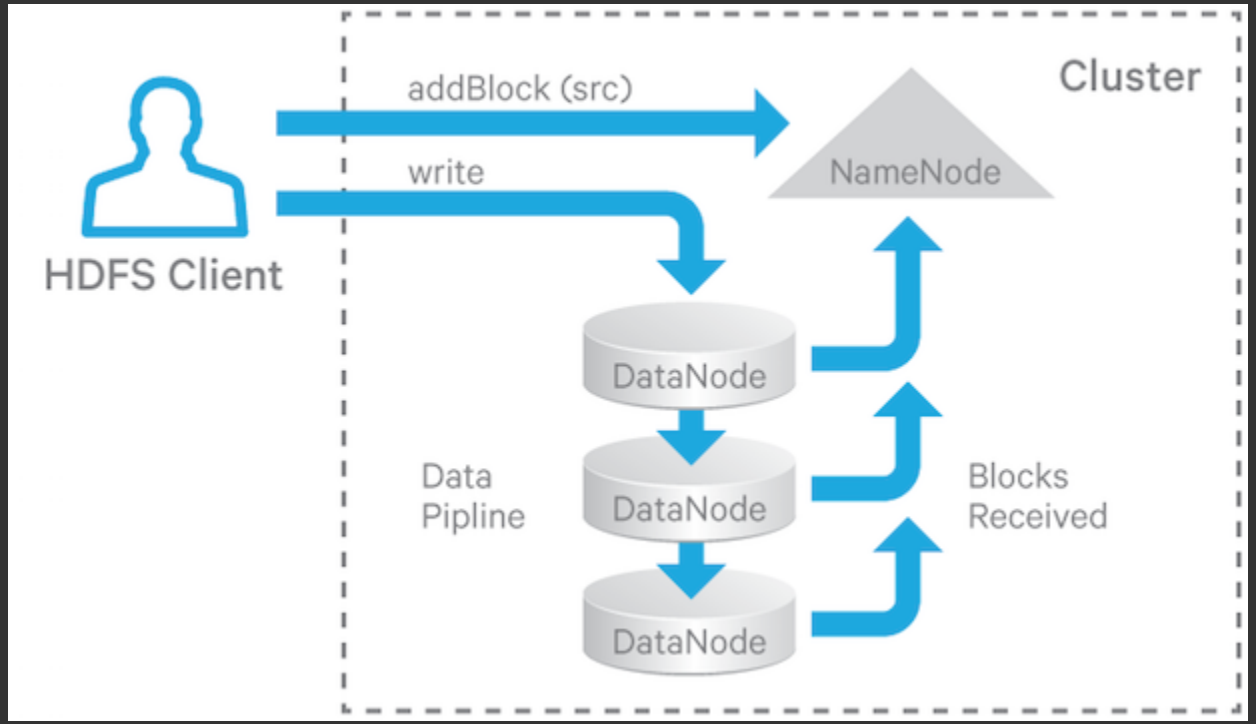
Architecture



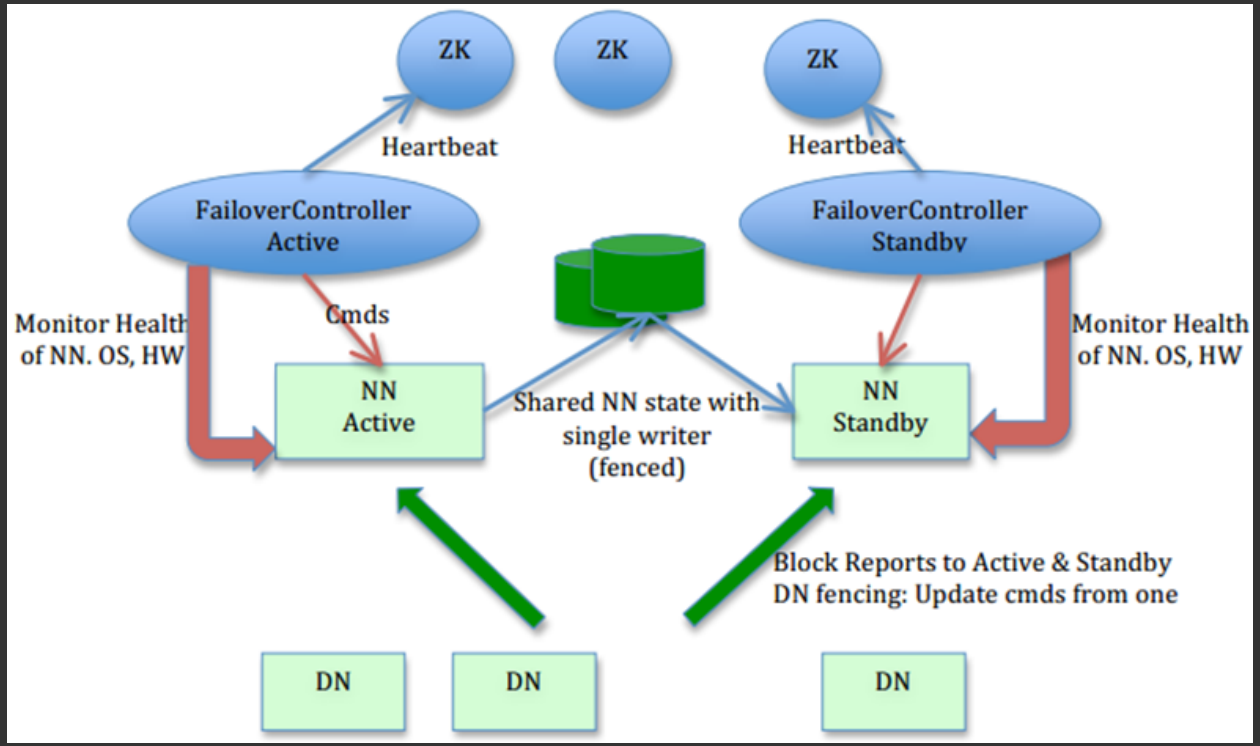
Replication



Pipeline


Cluster
addBlock (src)
write
NameNode
HDFS Client
DataNode
DataNode
DataNode
Data
Pipline
Blocks
Received

HA


ZK
ZK
ZK
Heartbeat
Heartbeat
FailoverController Active
FailoverController Standby
Monitor Health of NN. OS, HW
Cmds
Monitor Health of NN. OS, HW
NN Active
Shared NN state with single writer (fenced)
NN Standby
Block Reports to Active & Standby
DN fencing: Update cmds from one
DN
DN
DN

02 HDFS@toutiao

Overview

Capacity - 200+ PB

Performance

Throughput(avg)	-	95GB
RPC QPS(avg)	-	30000

Hardware

SSD / DISK

WARM / COLD

Challenge

Various application - mr, spark, hbase, hive

1. throughput & low latency requirement
2. IO activity affect each other

Challenge

Namenode memory pressure

1. namenode mixed gc - 30s
2. restart time - 1h

Challenge

Rpc request explode

1. hive server restart
2. resourcemanager restart
3. huge application stop

Challenge

IO isolation

1. datanode & nodemanager
2. application within hdfs

Challenge

Trouble shooting

1. no unique id through pipeline
2. client log, datanode log, namenode log

Challenge

Future

1. 1w+ node
2. multiple dc
3. diverse hardware
4. complex network condition

Challenge

Rpc request explode

1. hive server restart
2. resourcemanager restart
3. huge application end

03 What we done ?

HA

NNProxy - https://github.com/bytedance/nnproxy

Two name nodes for now

1. namenode1 - 84%
2. namenode2 - 16%

Data Compress Pipeline

namenode memory image -> hive -> mr -> hdfs

45PB

Heterogeneous Storage

36 cold-data nodes for compressed data

2 storage type - disk & ssd

Centralized Cache

Impala

Util - FastCopy

based on Facebook fastcopy

for hdfs 2.X

IO Control

IO throttle based on io util

04 Future HDFS @ toutiao

Target

1. block manage as a service
2. tracing system
3. centralized IO control system
4. quota

QA

Thanks


Broadview
Hadoop 2.X
HDFS源码剖析
徐鹏 著
中国工信出版集团
@博文视点Broadview
weibo.com/broadview2j