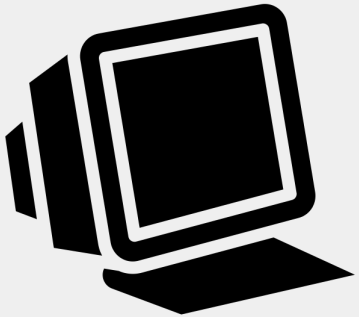
Hadoop 報告

BDSE18 第一組

李子顥 陳姿伶 陳彥伶 李謀勳 劉文裕

Hadoop 叢集規格:

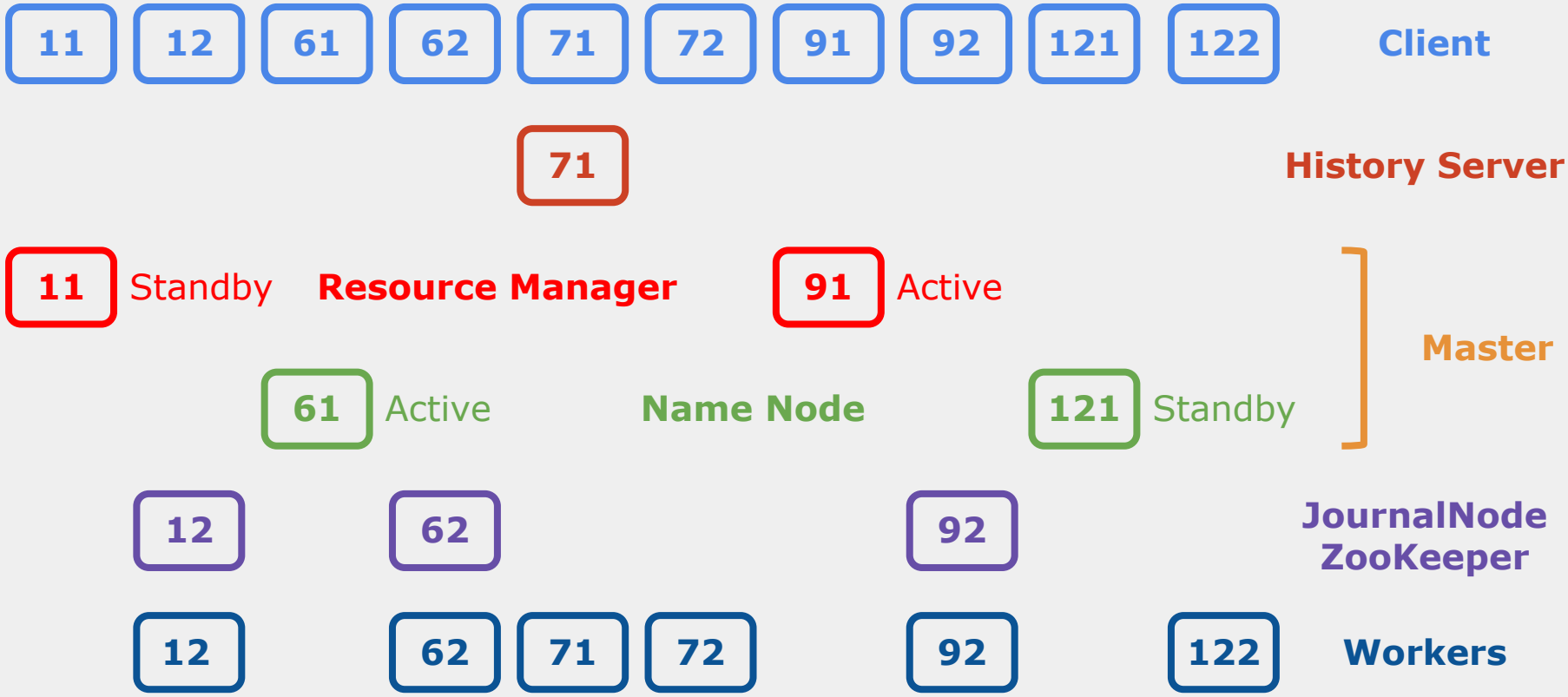


- 五台 實體主機:
  - 十台 虛擬主機:
    - 4 核心
    - 24GB RAM
    - 動態硬碟配置, 目前最高500GB

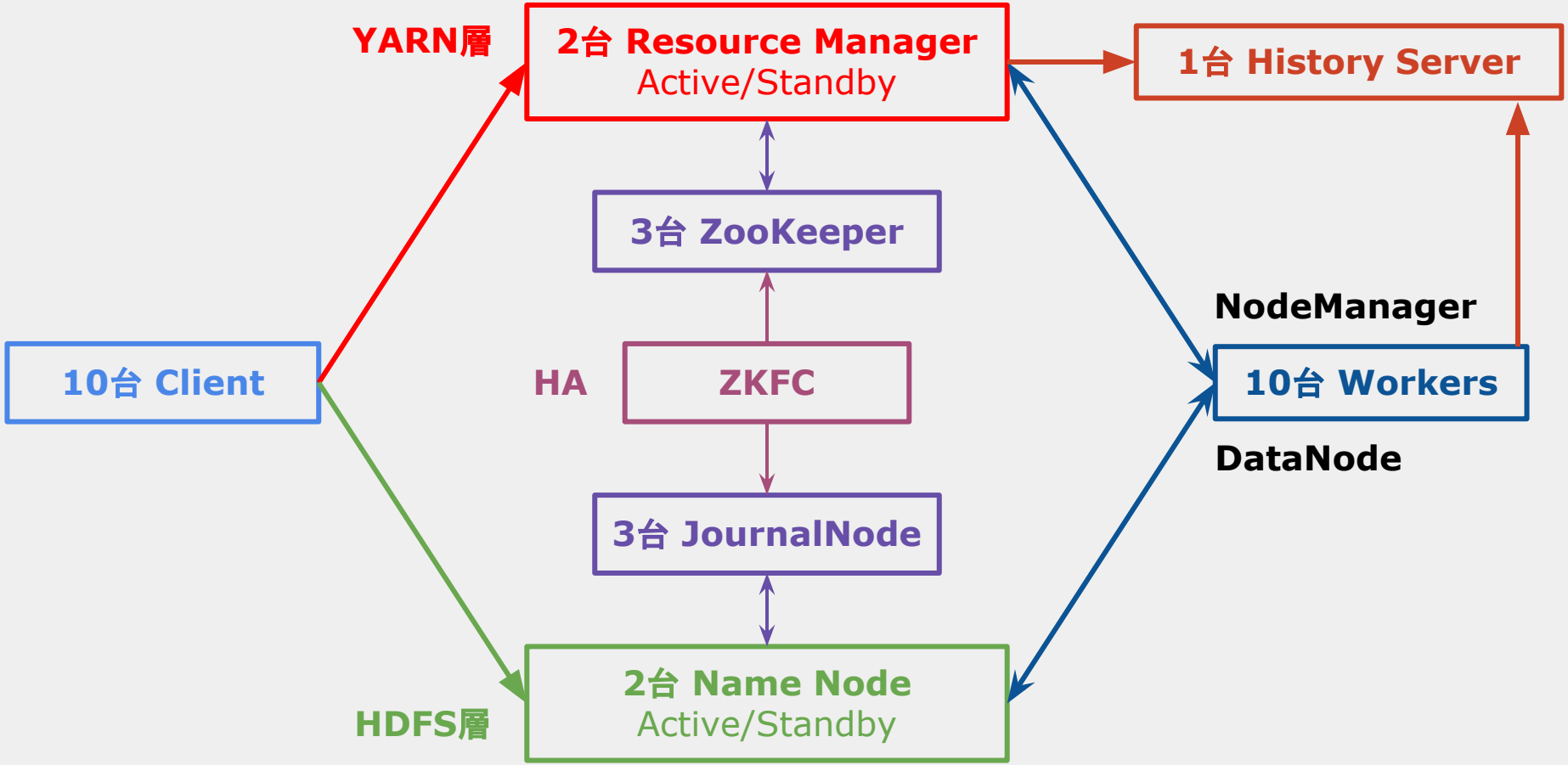
- 叢集:
  - 六台 Workers:
    - 24核心、144GB RAM、最高3TB

Hadoop 叢集架構配置

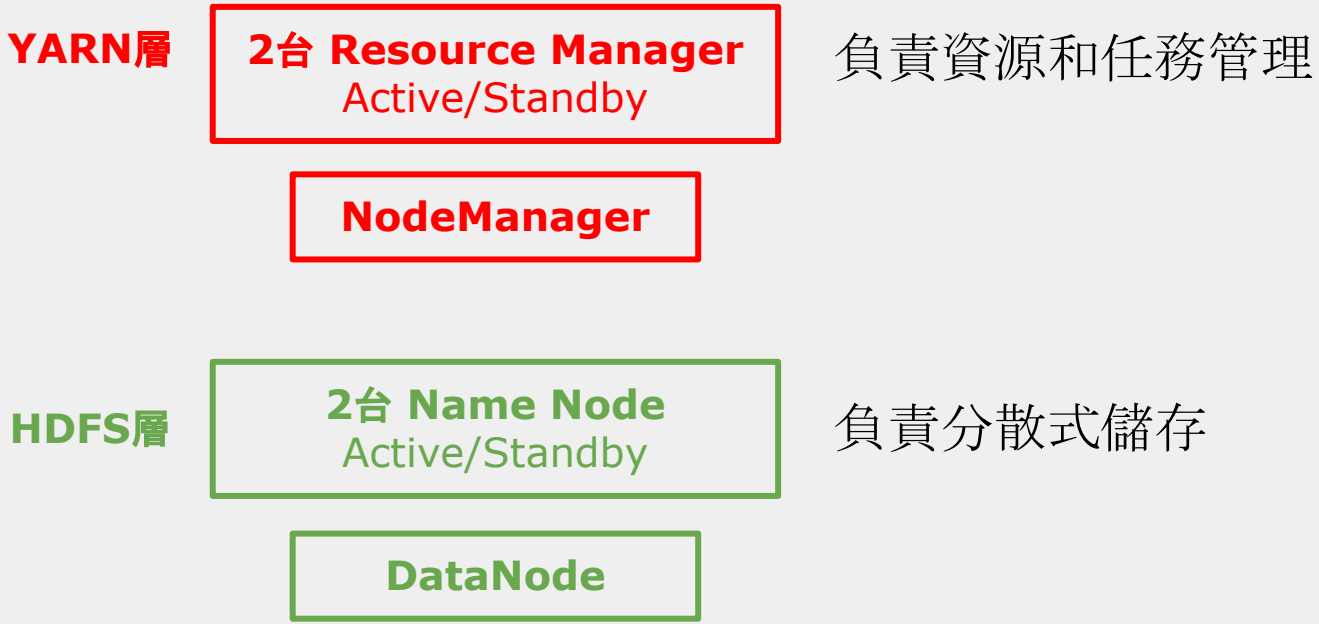
FQDN: bdse___.example.org



Hadoop HA叢集工作架構



Hadoop 節點簡介


YARN層
2台 Resource Manager
Active/Standby
NodeManager
負責資源和任務管理
HDFS層
2台 Name Node
Active/Standby
DataNode
負責分散式儲存

Hadoop 節點簡介

Map-Reduce　負責分散式計算

基於　YARN層 + HDFS層 + 1台 History Server

HA

3台 ZooKeeper

ZKFC

3台 JournalNode

高可用性, 監控狀態

切換Active/StandBy 的RM/NN

Demo 1. 程式執行(台股日成交，300mb，400萬筆)

Stock.ipynb

Code

```
[1]: # import modules
     import pandas as pd
     import numpy as np

[2]: from pyspark.sql import SparkSession

[3]: spark.sparkContext.appName

[3]: 'PySparkShell'

[4]: spark.conf.set("spark.sql.execution.arrow.pyspark.enabled", True)

[5]: import databricks.koalas as ks

[6]: ks.set_option("compute.default_index_type", "distributed")

[7]: %time perfdf = ks.read_csv("/user/stock/daily.csv")

     CPU times: user 2.28 ms, sys: 23.4 ms, total: 25.6 ms
     Wall time: 7.21 s

[8]: %time perfdf.shape

     CPU times: user 0 ns, sys: 7.84 ms, total: 7.84 ms
     Wall time: 751 ms
[8]: (4078037, 11)

[9]: perfdf
```

[9]:

	stock_id	stock_name	date	Volume	Volume_Cash	Open	High	Low	Close	Change	Order
0	1101	台泥	2010-01-04	8299290.0	2.835298e+08	34.00	34.40	33.90	34.30	0.30	2839.0
1	1101	台泥	2010-01-05	48059367.0	1.693662e+09	34.60	35.80	34.50	35.50	1.20	13190.0
2	1101	台泥	2010-01-06	36026739.0	1.298755e+09	35.50	36.40	35.40	36.25	0.75	9057.0
3	1101	台泥	2010-01-07	14531020.0	5.208832e+08	36.35	36.40	35.50	35.55	-0.70	4733.0
4	1101	台泥	2010-01-08	10671139.0	3.772868e+08	35.55	35.80	35.15	35.40	-0.15	3456.0
5	1101	台泥	2010-01-11	18206405.0	6.380753e+08	35.40	35.80	34.70	34.75	-0.65	4963.0
6	1101	台泥	2010-01-12	15970864.0	5.536289e+08	34.75	35.25	34.30	34.55	-0.20	4895.0
7	1101	台泥	2010-01-13	23801821.0	8.094711e+08	34.00	34.25	33.90	34.00	-0.55	5839.0
8	1101	台泥	2010-01-14	8192369.0	2.803680e+08	34.50	34.50	34.10	34.20	0.20	2673.0

Demo 2. TroubleShooting 實際案例

```
hadoop@bdse91:~$ start-yarn.sh
start-yarn.sh: command not found
hadoop@bdse91:~$ echo $PATH
/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/games:/usr/lo
cal/games:/snap/bin:/usr/local/hadoop/bin:/usr/local/hadoop/sbin1
hadoop@bdse91:~$ exit
logout
ubuntu@bdse91:~$ su - hadoop
Password:
hadoop@bdse91:~$ echo $PATH
/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/games:/usr/lo
cal/games:/snap/bin:/usr/local/hadoop/bin:/usr/local/hadoop/sbin
hadoop@bdse91:~$ start-yarn.sh
Starting resourcemanager
WARNING: /usr/local/hadoop/logs does not exist. Creating.
Starting nodemanagers
bdse72.example.org: Warning: Permanently added 'bdse72.example.org' (ECDSA) to
the list of known hosts.
bdse71.example.org: Warning: Permanently added 'bdse71.example.org' (ECDSA) to
the list of known hosts.
bdse122.example.org: Warning: Permanently added 'bdse122.example.org,192.168.32
.122' (ECDSA) to the list of known hosts.
bdse12.example.org: Warning: Permanently added 'bdse12.example.org,192.168.32.1
2' (ECDSA) to the list of known hosts.
bdse62.example.org: Warning: Permanently added 'bdse62.example.org,192.168.32.6
2' (ECDSA) to the list of known hosts.
bdse92.example.org: Warning: Permanently added 'bdse92.example.org,192.168.32.9
2' (ECDSA) to the list of known hosts.
hadoop@bdse91:~$
```

Demo 2. TroubleShooting查log

1. **除錯 (查log)**

```
cd $HADOOP_HOME
cd logs/
ls -l
```

```
bdse[NodeManager].example.org:8042/logs
```

2. **檢查所有.xml檔案**