

# 安裝設定補充說明

王耀聰 陳威宇

jazz@nchc.org.tw

wuae@nchc.org.tw

國家高速網路與計算中心(NCHC)



# Hadoop Package Topology

## 資料夾

## 說明

bin /	各執行檔：如 start-all.sh 、 stop-all.sh 、 hadoop
conf /	預設的設定檔目錄：設定環境變數 hadoop-env.sh 、各項參數 hadoop-site.conf 、工作節點 slaves 。 ( 可更改路徑 )
docs /	Hadoop API 與說明文件 ( html & PDF)
contrib /	額外有用的功能套件，如：eclipse的擴充外掛、Streaming 函式庫 。
lib /	開發 hadoop 專案或編譯 hadoop 程式所需要的所有函式庫，如：jetty 、 kfs 。但主要的hadoop函式庫於hadoop_home
src /	Hadoop 的原始碼。
build /	開發Hadoop 編譯後的資料夾。需搭配 ant 程式與build.xml
logs /	預設的日誌檔所在目錄。 ( 可更改路徑 )

# 環境參數：hadoop-env.sh

- 設定Linux系統執行Hadoop的環境參數
  - export xxx=kkk
    - 將kkk這個值匯入到xxx參數中
  - # string…
    - 註解，通常用來描述下一行的動作內容

```
# The java implementation to use. Required.  
export JAVA_HOME=/usr/lib/jvm/java-6-sun  
export HADOOP_HOME=/opt/hadoop  
export HADOOP_LOG_DIR=$HADOOP_HOME/logs  
export HADOOP_SLAVES=$HADOOP_HOME/conf/slaves  
.....
```

# 設定檔： core-site.xml (0.20)

## <configuration>

```
<property>
  <name> fs.default.name</name>
  <value> hdfs://localhost:9000/</value>
  <description> ... </description>
</property>
```

```
<property>
  <name> hadoop.tmp.dir </name>
  <value> /tmp/hadoop/hadoop-
    ${user.name} </value>
  <description> ... </description>
</property>
```

## <configuration>

詳細 hadoop core 參數，

請參閱 <http://hadoop.apache.org/common/docs/current/core-default.html>

以及 \$HADOOP\_HOME/src/core/core-default.xml

# 設定檔： mapred-site.xml (0.20)

## <configuration>

```
<property>
  <name> mapred.job.tracker</name>
  <value> localhost:9001</value>
  <description>... </description>
</property>
```

```
<property>
  <name> mapred.map.tasks</name>
  <value> 1</value>
  <description> ... </description>
</property>
```

## <property>

```
  <name> mapred.reduce.tasks</name>
  <value> 1</value>
  <description> ... </description>
</property>
</configuration>
```

詳細 hadoop mapreduce 參數，  
請參閱 <http://hadoop.apache.org/common/docs/current/mapred-default.html>  
以及 \$HADOOP\_HOME/src/mapred/mapred-default.xml



# 設定檔： hdfs-site.xml (0.20)

<configuration>

```
<property>
  <name> dfs.replication </name>
  <value> 3</value>
  <description>... </description>
</property>
```

```
<property>
  <name> dfs.permissions </name>
  <value> false </value>
  <description> ... </description>
</property>
```

</configuration>

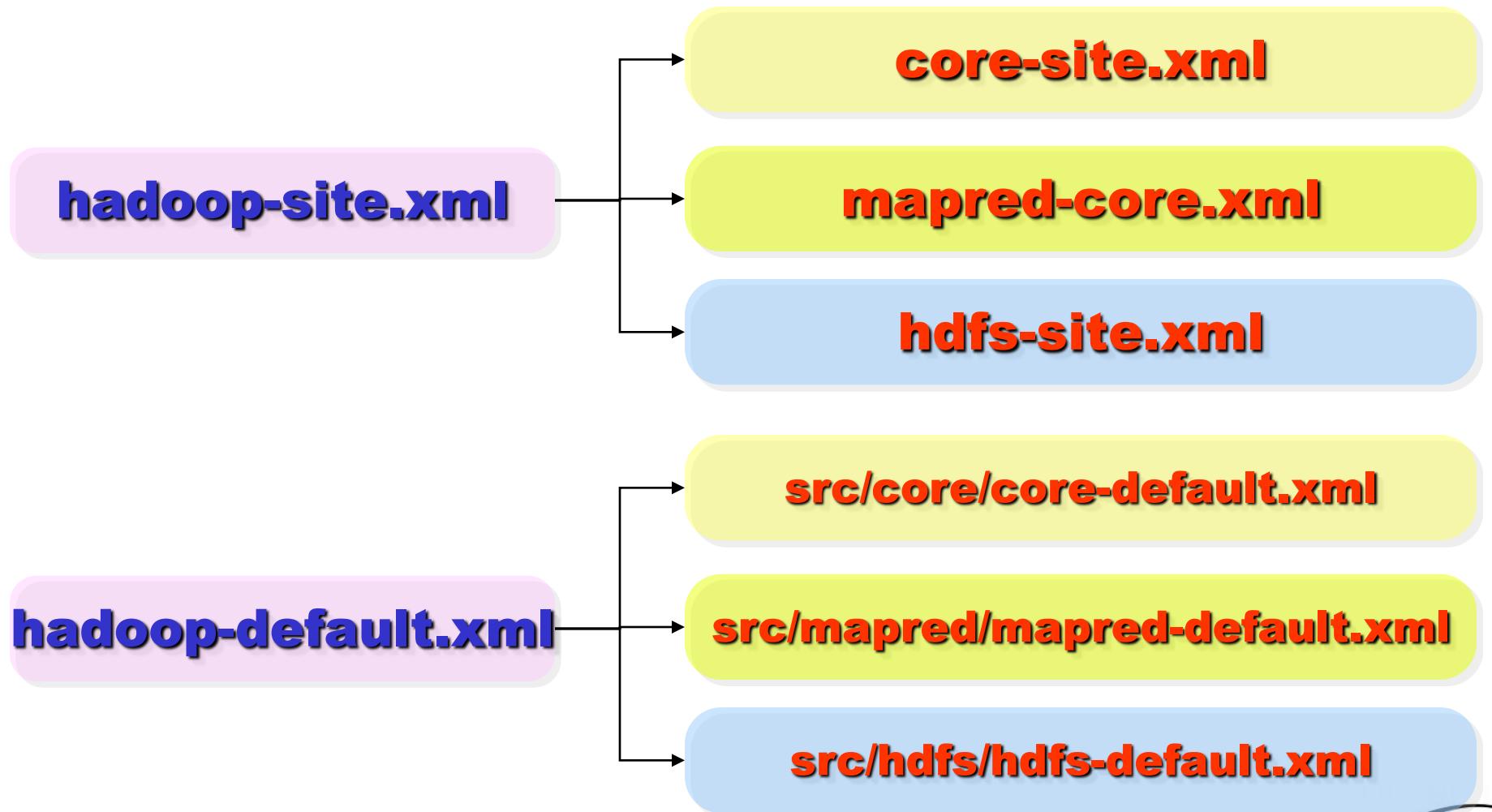
詳細 hadoop hdfs 參數，

請參閱 <http://hadoop.apache.org/common/docs/current/hdfs-default.html>

以及 \$HADOOP\_HOME/src/hdfs/hdfs-default.xml



# Hadoop 0.18 到 0.20 的轉變



# 設定檔：hadoop-site.xml (0.18)

## <configuration>

```
<property>
  <name> fs.default.name</name>
  <value> hdfs://localhost:9000</value>
  <description> ... </description>
</property>

<property>
  <name> mapred.job.tracker</name>
  <value> localhost:9001</value>
  <description>... </description>
</property>

<property>
  <name> hadoop.tmp.dir </name>
  <value> /tmp/hadoop/hadoop-
    ${user.name} </value>
  <description> </description>
</property>
```

## <property>

```
<name> mapred.map.tasks</name>
<value> 1</value>
<description> define mapred.map tasks to
  be number of slave hosts
</description>

</property>

<property>
  <name> mapred.reduce.tasks</name>
  <value> 1</value>
  <description> define mapred.reduce tasks
    to be number of slave hosts
  </description>
</property>

<property>
  <name> dfs.replication</name>
  <value> 3</value>
</property>

</configuration>
```

# 設定檔：hadoop-default.xml (0.18)

- Hadoop預設參數
  - 沒在hadoop.site.xml設定的話就會用此檔案的值
  - 更多的介紹參數：  
[http://hadoop.apache.org/core/docs/current/cluster\\_setup.html#Configuring+the+Hadoop+Daemons](http://hadoop.apache.org/core/docs/current/cluster_setup.html#Configuring+the+Hadoop+Daemons)

# 設定檔： slaves

- 紿 start-all.sh , stop-all.sh 用
- 被此檔紀錄到的節點就會附有兩個身份：  
datanode & tasktracker
- 一行一個hostname 或 ip

```
192.168.1.1  
....  
192.168.1.100  
Pc101  
....  
Pc152  
....
```

# 設定檔：masters

- 紿 start-\* .sh , stop-\* .sh 用
- 會被設定成 secondary namenode
- 可多個

192.168.1.1

....

Pc101

....

描述名稱	設定名稱	所在檔案
JAVA安裝目錄	JAVA_HOME	hadoop-env.sh
HADOOP家目錄	HADOOP_HOME	hadoop-env.sh
設定檔目錄	HADOOP_CONF_DIR	hadoop-env.sh
日誌檔產生目錄	HADOOP_LOG_DIR	hadoop-env.sh
HADOOP工作目錄	hadoop.tmp.dir	hadoop-site.xml
JobTracker	mapred.job.tracker	hadoop-site.xml
Namenode	fs.default.name	hadoop-site.xml
TaskTracker	(hostname)	slaves
Datanode	(hostname)	slaves
第二Namenode	(hostname)	masters
其他設定值	詳可見hadoop-default.xml	hadoop-site.xml

# 控制 Hadoop 的指令

- 格式化
  - \$ bin/hadoop namenode -format
- 全部開始(透過 SSH )      • 全部結束(透過 SSH )
  - \$ bin/start-all.sh
  - \$ bin/start-dfs.sh
  - \$ bin/start-mapred.sh
  - \$ bin/stop-all.sh
  - \$ bin/stop-dfs.sh
  - \$ bin/stop-mapred.sh
- 獨立啟動/關閉(不會透過 SSH )
  - \$ bin/hadoop-daemon.sh [start/stop] namenode
  - \$ bin/hadoop-daemon.sh [start/stop] secondarynamenode
  - \$ bin/hadoop-daemon.sh [start/stop] datanode
  - \$ bin/hadoop-daemon.sh [start/stop] jobtracker
  - \$ bin/hadoop-daemon.sh [start/stop] tasktracker

# Hadoop 的操作與運算指令

- 使用hadoop檔案系統指令
  - \$ bin/hadoop  $\Delta$  fs  $\Delta$  –Instruction  $\Delta$  ...
- 使用hadoop運算功能
  - \$ bin/hadoop  $\Delta$  jar  $\Delta$  XXX.jar  $\Delta$  Main\_Function  $\Delta$  ...

# Hadoop 使用者指令

\$ bin/hadoop  $\Delta$  指令  $\Delta$  選項  $\Delta$  參數  $\Delta$  ....

指令	用途	舉例
<b>fs</b>	對檔案系統進行操作	hadoop $\Delta$ <b>fs</b> $\Delta$ -put $\Delta$ in $\Delta$ input
<b>jar</b>	啟動運算功能	hadoop $\Delta$ <b>jar</b> $\Delta$ example.jar $\Delta$ wc $\Delta$ in $\Delta$ out
<b>archive</b>	封裝hdfs上的資料	hadoop $\Delta$ <b>archive</b> $\Delta$ foo.har $\Delta$ /dir $\Delta$ /user/hadoop
<b>distcp</b>	用於叢集間資料傳輸	hadoop $\Delta$ <b>distcp</b> $\Delta$ hdfs://nn1:9000/aa $\Delta$ hdfs://nn2:9000/aa
<b>fsck</b>	hdfs系統檢查工具	hadoop $\Delta$ <b>fsck</b> $\Delta$ /aa $\Delta$ -files $\Delta$ -blocks $\Delta$ -locations
<b>job</b>	操作正運算中的程序	hadoop $\Delta$ <b>job</b> $\Delta$ -kill $\Delta$ jobID
<b>version</b>	顯示版本	hadoop $\Delta$ <b>version</b>

# Hadoop 管理者指令

\$ bin/hadoop  $\Delta$ 指令  $\Delta$ 選項  $\Delta$ 參數  $\Delta$ ....

指令	用途	舉例
<b>balancer</b>	平衡hdfs覆載量	hadoop $\Delta$ <b>balancer</b>
<b>dfsadmin</b>	配額、安全模式 等管理員操作	hadoop $\Delta$ <b>dfsadmin</b> $\Delta$ -setQuota $\Delta$ 3 $\Delta$ /user1/
<b>namenode</b>	名稱節點操作	hadoop $\Delta$ <b>namenode</b> $\Delta$ -format

\$ bin/hadoop  $\Delta$ 指令

datanode	成為資料節點	hadoop $\Delta$ datanode
jobtracker	成為工作分派者	hadoop $\Delta$ jobtracker
tasktracker	成為工作執行者	hadoop $\Delta$ tasktracker