Planificar el clúster Hadoop

Iván González

Agenda

- Consideraciones iniciales
- Elegir el hardware adecuado
- Red de comunicaciones
- Configuración de los nodos

Parte del contenido de esta presentación ha sido extraído de la web de Cloudera http://www.cloudera.com

Consideraciones iniciales

- Es recomendable empezar con un clúster pequeño para familiarizarse con la instalación y administración de Hadoop
  - Por ejemplo 4 o 5 nodos
- A medida que sea necesario, se puede ir incrementando el número de nodos
- ¿Cuándo es necesario aumentar el clúster?
  - Si se necesita más potencia de computo
  - Si aumenta la cantidad de datos a almacenar
  - Si aumenta la cantidad de memoria requerida por los procesos a ejecutar

Consideraciones iniciales

- Basar el crecimiento del clúster en función del almacenamiento puede ser una buena metodología
- Por ejemplo, supongamos el siguiente escenario
  - Los datos aumentan en 4 TB cada semana
  - HDFS está configurado por defecto para replicar 3 veces
  - Esto supone que cada semana necesitamos 12 TB
    - Se debe añadir un poco más, por ejemplo un 25%, para overhead
      - Ficheros locales temporales, etc.
    - En total 15 TB a la semana
  - Si asumimos servidores con 12 x 4 TB discos, necesitamos una nueva máquina cada 3 semanas
    - 2 años de datos suponen 1,5 PB lo que suponen aprox. 30 máquinas

Elegir el hardware adecuado

- Los nodos de un clúster Hadoop se pueden clasificar como "esclavos" o "maestros"
- Los nodos esclavos ejecutan un DataNode y un NodeManager
- Los nodos maestro ejecutan el NameNode, el Secondary NameNode o el ResourceManager
  - En clúster pequeños, el NameNode y el ResourceManager pueden estar en el mismo nodo
  - Incluso se podría ejecutar el Secondary NameNode en el mismo nodo
    - Pero en este caso es importante tener un copia adicional de los meta-datos del NameNode en otra máquina

Elegir el hardware adecuado – Nodos esclavo

- Las recomendaciones típicas para los nodos esclavo son:
  - Conf. media: Alta densidad de almacenamiento, 1 Gb Ethernet
    - 12 x 3 TB discos SATA en JBOD (Just a Buch Of Disks)
    - 2 x 6-core 2,9 GHz CPUs
    - 64 GB RAM
    - 2x1 Ethernet
  - Conf. alta: Mucha memoria, uso intensivo del almacenamiento, 10 Gb Ethernet
    - 24 x 1 TB discos SAS/SSD en JBOD
    - 2 x 6-core 2,9 GHz CPUs
    - 96 GB RAM
    - 1x10 Gb Ethernet

Elegir el hardware adecuado – Nodos esclavo - CPU

- Las CPUs de 6 cores son muy comunes actualmente
  - Los servidores con dos procesadores son lo más adecuados
    - Dispondremos de 12 cores físicos por nodo
    - Con HyperThreading (Intel) serían 24 cores lógicos
  - A más cores más memoria y almacenamiento son necesarios
- El HyperThreading (Intel) se recomienda activarlo
- Los procesos Hadoop no suelen estar limitados por la capacidad de procesamiento
  - Típicamente la limitación suele venir de los discos o de la red
  - Por tanto no es necesario emplear CPUs "tope de gama"

Elegir el hardware adecuado – Nodos esclavo - RAM

- La configuración de los nodos esclavo determina el número de tareas Map y Reduce que pueden ejecutarse simultáneamente
- Cada tarea Map y Reduce suele emplear entre 2 y 4 GB de RAM
- Es importante evitar el uso de memoria virtual (swap)
  - Hay que disponer de memoria suficiente para todas las tareas MR, más un overhead para los demonios que se ejecutan en el nodo: DataNode, NodeManager, etc.
- Una regla que se puede aplicar inicialmente
  - N° total de tareas = 1,5 x n° de cores físicos de CPU
  - No tiene porque ser adecuada para todos los clústeres

Elegir el hardware adecuado – Nodos esclavo - Discos

- La arquitectura de Hadoop afecta a los requerimientos de almacenamiento
  - Por defecto HDFS replica los datos 3 veces
  - Se almacenan datos temporales que requieren típicamente entre el 20%-30% del almacenamiento total
- En general, cuantos más discos mejor
  - En práctica suelen usarse entre 4 y 24 discos por nodo
- Se recomienda usar discos 3,5" ya que son igual de rápidos, más baratos y con mayor capacidad que los discos de 2,5"
  - SATA 7200 RPM es suficiente

Elegir el hardware adecuado – Nodos esclavo - Discos

- Usar 8 x 1,5 TB es probablemente mejor que 6 x 2 TB
  - Es más probable que cada tarea acceda a un disco diferente
- Un máximo adecuado pueden ser 36 TB por nodo
  - Más cantidad supondrá un aumento importante del tráfico de red si un nodo "muere" y es necesario re-replicar los datos

Elegir el hardware adecuado – Nodos esclavo - Discos

- ¿Por qué no se usa RAID (Redundant Array of Inexpensive Disks)?
  - HDFS ya incluye redundancia al replicar los bloques entre diferentes nodos
  - RAID 0 (Data stripping) es la opción RAID que ofrece mejor rendimiento
    - Sin embargo, es más lento que JBOD cuando se emplea HDFS
    - El rendimiento del RAID 0 está limitado por el disco más lento
    - Las operaciones de disco en JBOD son independientes, por lo que la velocidad media es superior a la del disco más lento
    - Un test ejecutado por Yahoo mostraba que el rendimiento de JBOD es entre un 30% y 50% más rápido que RAID 0, dependiendo de las operaciones realizadas.

Elegir el hardware adecuado – Nodos esclavo

- ¿Y que pasa si virtualizamos los nodos?
  - La virtualización supondrá una penalización en rendimiento y confiabilidad, incluyendo
    - Contención de red
    - Se suelen emplear discos remotos y configurados como un único volumen, aunque el disco sea local
    - No hay modo de garantizar el alojamiento de los nodos en "diferentes racks"
    - Es posible que las tres réplicas de los datos se encuentren en la misma máquina física
  - Siempre que sea posible se recomienda usar hardware físico dedicado para el clúster
  - Es razonable el uso de virtualización para pruebas de concepto de clústeres o cuando no sea posible el uso de hardware dedicado (restricciones del centro de datos, por ejemplo)

Elegir el hardware adecuado – Nodos esclavo

- ¿Y que pasa si usamos blades?
  - En general no se recomiendan
  - El fallo de un chasis de blades puede suponer el fallo de varios nodos
  - Los blades suelen tener poca capacidad de almacenamiento
  - La interconexión de red de los blades dentro del chasis podría ser un cuello de botella

Elegir el hardware adecuado – Nodos maestro

- Es recomendable usar hardware "bueno" (first class) en vez de hardware "commodity"
- Doble fuente de alimentación
- Tarjetas de red de doble puerto
  - Se combinan (bonded) los puertos para mitigar posibles fallos
- Discos en RAID
- Memoria RAM razonable
  - 24 GB para clústeres de 20 nodos o menos
  - 48 GB para clústeres de hasta 300 nodos
  - 96 GB para clústeres más grandes

Elegir el hardware adecuado – Fallo de un nodo

- Se asume que un nodo esclavo fallará en algún momento
  - Con esta idea se desarrolló Hadoop
  - El NameNode automáticamente replicará los bloques del nodo, manteniendo las 3 réplicas.
  - El ResourcerManager automáticamente reasignará las tareas a otros nodos
- El fallo de un nodo maestro si es un problema importante si no se ha configurado el clúster en Alta Disponibilidad (AD)
  - Si cae el NameNode los datos del clúster son inaccesibles
  - Si cae el ResourceManager no se pueden ejecutar trabajos en el clúster
    - Todos los trabajos que se estén ejecutando fallarán
  - Configurar el NameNode y el ResourceManager en AD es lo recomendado para sistemas de producción

Red de comunicaciones

- Hadoop hace un uso intensivo de la red
  - Todos los nodos se comunican entre si
- Los nodos deberían estar conectados mínimo a 1 Gbps
- Considerar 10 Gbps en los siguientes casos:
  - Clústeres con una gran cantidad de datos
  - Clústeres donde los trabajos MapReduce típicos generan gran cantidad de datos intermedios
- Es recomendable usar switches dedicados para el clúster
  - Los nodos de un mismo rack deben conectarse a switches top-of-rack y los racks a través de core switches de 10 Gbps o más rápidos
  - Evitar sobrecarga en los switches
- Considerar combinar puertos (bonded) para mitigar fallos
- Considerar switches redundantes

Red de comunicaciones

- Resolución de nombres
  - Cuando se configura Hadoop se requiere identificar algunos nodos en los ficheros de configuración de Hadoop
  - Usar nombres en vez de IPs para identificar los nodos
  - Usar DNS o el fichero */etc/hosts*
  - Hadoop usa la resolución de nombres directa e inversa
    - En caso de no resolver, podrían ocurrir problemas
    - Por ejemplo, si los esclavos se encuentran en una red interna no será posible resolver los nombres desde un cliente externo
- Deshabilitar el firewall dentro del clúster si es posible (red interna) o permitir acceso total entre máquinas del clúster
  - Evita problemas de acceso a los nodos
  - Hadoop emplea muchos puertos de red y es difícil tenerlo todo controlado

Configuración de nodos – Sistema Operativo

- Elige el sistema operativo con el que estés más cómodo en la parte de administración
- CentOS
  - Se considera una versión más orientada a servidor
  - Muy conservadora en cuanto a las versiones de paquetes
  - Muy utilizada en producción
- RedHat
  - Igual que CentOS pero incluye soporte de pago

Configuración de nodos – Sistema Operativo

- Elige el sistema operativo con el que estés más cómodo en la parte de administración
- CentOS
  - Se considera una versión más
  - Muy conservadora en cuanto
  - Muy utilizada en producción
- RedHat
  - Igual que CentOS pero incluye soporte de pago

En producción es típico mezclar RedHat y CentOS

- RedHat en nodos maestro
- CentOS en nodos esclavo

Configuración de nodos – Sistema Operativo

- Ubuntu
  - Muy popular
  - Basada en Debían
  - Disponibles versiones de servidor y escritorio
  - Es recomendable usar la variante LTS (Long Term Support)
- SuSE
  - Del estilo de RedHat/CentOS, aunque con diferencias importantes en la organización de directorios y ficheros del sistema, y en la gestión de paquetes
  - Es muy popular también en entornos de producción

Configuración de nodos – Configuración del sistema

- No usar Linux LVM (Logical Volume Manager) para hacer que todos los discos aparezcan como uno solo
  - Al igual que con RAID 0 está limitado al disco más lento
- Chequear las opciones de la BIOS (Basic Input/Ouput System)
  - Puede que no estén configuradas para máximo rendimiento
  - Por ejemplo, si se usan discos SATA habría que deshabilitar la emulación IDE
- Hacer un test de disco mediante *hdparm –t*
  - **Ejemplo:** `hdparm -t /dev/sda1`
  - Deberías ver velocidades cercanas a 70 MB/s o más. Menos que eso indicaría problemas

Configuración de nodos – Configuración del sistema

- Hadoop no requiere un particionado específico de los discos
  - Usa cualquiera que tenga sentido
- Montar los discos con la opción *noatime*
  - Evita que cada vez que se acceda a un fichero o carpeta se realice una escritura en disco para actualizar la fecha de acceso
- Estructura de directorios típica para montar los discos de datos:

```
/data/<n>/dfs/nn
/data/<n>/dfs/dn
/data/<n>/dfs/snn
/data/<n>/mapred/local
```

- Evitar o reducir el swap del sistema
  - Asignar 0 o 5 a *vm.swappinnes* en */etc/sysctl.conf*

Configuración de nodos – Sistema de ficheros

- Se suele recomendar ext3 o ext4
  - ext4 se emplea comúnmente en nuevos clústeres
- XFS es otra buena opción que aporta algunos beneficios durante la puesta en marcha del clúster
  - Formatea en 0 segundos frente varios minutos que requiere cada disco en ext3/ext4

Configuración de nodos – Parámetros del SO

- Incrementar el *nofile ulimit* para los usuarios *mapred* y *hdfs* al menos a 32K
  - En */etc/security/limits.conf*
- Deshabilitar IPv6
- Deshabilitar SELinux si es posible
  - Incurre en un 7-10% de penalización en rendimiento
  - La configuración no es sencilla
- Instalar y configurar el demonio ntp
  - Garantiza que todos los nodos están sincronizados
  - Importante para varias aplicaciones del ecosistema Hadoop
  - Útil si se usan los logs para identificar problemas

Configuración de nodos – Máquina Virtual Java

- Es recomendable usar la distribución oficial de Oracle JDK
  - Hadoop es complejo y es posible que aparezcan bugs en otras distribuciones
- Chequea la web de Hadoop para ver que versiones de Java están soportadas
  - Versión 2.7 de Hadoop requiere Java 7
  - Versiones anteriores requieren Java 6

Configuración de nodos – Clústeres grandes

- Cada nodo del clúster requiere sus propios ficheros de configuración
- Administrar pequeños clústeres es relativamente sencillo
  - Accede a cada máquina y realiza los cambios necesarios "manualmente"
- A medida que el clúster crece, se hace más complejo
- Existen herramientas de administración que permiten administrar varias máquinas a la vez
  - Actualizar ficheros, reiniciar demonios o incluso reiniciar máquinas automáticamente cuando sea necesario

Configuración de nodos – Clústeres grandes

- Recomendaciones
  - Usar herramientas de administración
  - Empezar a usarlas incluso con pequeños clústeres
  - Por ejemplo, Cloudera Manager
    - La edición gratuita soporta clústeres de hasta 50 nodos
- Herramientas opensource de administración de configuración más populares son Puppet, Chef, Ansible, SaltStack, Fabric…
  - Existen otras, también de pago