Data Science

Data Life Cycle

Backup

- Computación Avanzada y e-Ciencia – IFCA
- Ibán Cabrillo Bartolomé
- Santander Enero del 2020
-

UNIVERSIDAD DE CANTABRIA

Instituto de Física de Cantabria

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

Schema

- Backup
  - Models
  - Storage Status
  - Process
  - Data Management
  - Backup process
  - Care about
  - Backup Policies
  - Bacula backup software
    - Architecture
    - Installation
    - Configuration
    - Examples
      - Tape volumen creation
      - Add client node to bacula

Summary

- Physical Storage
- Network Storage
- File Systems & Network File Systems
- Data Management
- Backup

Backup I

- Copying and archiving of computer data so it may be used to *restore* the original after a data loss event
  - Recover data after its loss, be it by data deletion or corruption
  - Recover data from an earlier time, according to a user-defined data retention policy
- Models
  - Unstructured
    - Stack of floppy disks or CD-R/DVD-R media with minimal information about what was backed up and when
  - System Imaging/Full Backup
    - Complete system images from one or more specific points in time
  - Diferential
    - Saves the data since the last full backup
    - First necessary to perform a *full* backup
  - Incremental
    - Store backups from more points in time by organizing the data into increments of change between points in time
    - Eliminates the need to store duplicate copies of unchanged data
    - First necessary to perform a *full* backup

Backup II

- Reverse delta
  - recent mirror of the source data and a series of differences between the mirror in its current state and its previous states
  - First necessary to perform a *full* backup
  - rdiff-backup
- Continuous data protection
  - disk mirroring in that it enables a roll-back of the log and thus restoration of old image of data
- Snapshots
  - A snapshot is an instantaneous function of some storage systems that presents a copy of the file system as if it were frozen at a specific point in time
  - Subsequent snapshots copy the changed data only, and use a system of pointers to reference the initial snapshot
  - NTFS, access to snapshots is provided by the Volume Shadow-copying Service
  - ZFS, LVM, GPFS, BrtFS

Backup III

- Storage Media
  - HDDs
  - Optical Devices
  - Tapes
  - Remote Backup Service
- Attend to Storage media
  - On-line
    - Disk storage
    - restore in very sort time
    - Expensive
    - On-line storage is quite vulnerable to being deleted or overwritten
  - Near-line
    - less accessible and less expensive than on-line storage
    - tape library with restore times a few minutes
  - Off-line
    - direct human action in order to make access to the storage media physically possible
    - the data is not accessible via any computer except during limited periods
    - largely immune to a whole class of on-line backup failure modes
    - Manual Tape/DVD charger, external disk, etc

Backup IV

  - Off-Site
    - Protects against facilyty disaster
    - Can be on-line, nearl-line, remotely accessible
  - Disaster recovery (RD-Center)
    - Big infraestructures
- Data Management
  - Deciding what to back up at any given time is a harder process than it seems
  - backing up too much redundant data, the data repository will fill up too quickly
  - Backing up an insufficient amount of data can eventually lead to the loss of critical information.
  - Files
    - the simplest and most common way to perform a backup
  - File System Dump
    - Disk Imagin, unix dd, ZFS or XFS dump
  - Metadata
    - Boot Sector, Partition layout, file metadata, acls

Backup V

- Data Optimization
  - Compression
  - Deduplication
    - multiple similar systems are backed up to the same destination storage device
  - Encryption
  - Staging
    - Backup jobs are copied to a staging disk before being copied to tape
- Backup Proccess
  - backups need to be updated
  - Goals
    - Recovery Point Objetive (RPO)
      - The most desirable RPO would be the point just prior to the data loss event
      - requires increasing the frequency of synchronization between the source data and the backup repository
    - Recovery Time Objetive (RTO)
      - The amount of time elapsed between disaster and restoration
    - Data Security
      - Data does not compromise the original (Encription, WORM)

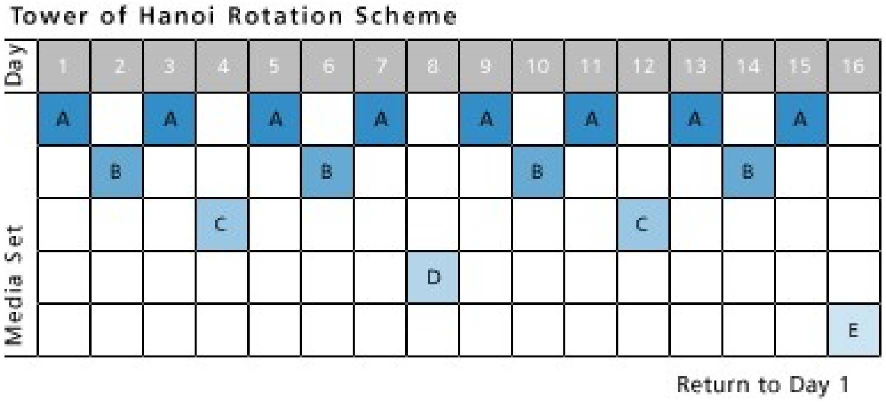
Backup VI

- To be Care
  - Window
    - Period of time when backups are permitted to run on a system
  - Performance
    - Backup schemes have some performance impact on the system being backed up
  - Network Bandwidth
    - Distributed backups systems can be affected
- Best Practises
  - Scheduling
    - The backup process uses CPU, memory, and network resources, along with disk I/O operations
  - Locations
    - Ensure that your backup media are in a different physical location from the main site
  - Check
    - Be sure that the backups work

Backup VII

- Backup Routines
  - Fisrt in Firts out
    - Backup scheme saves new or modified files onto the oldest media in the set
    - Performing a daily backup onto a set of 14 media, the backup depth would be 14 days
    - Used to keep the longest possible tail of daily backups
    - Used when archived backups are not as importan
    - Useful when data before the rotation period is irrelevant
  - Grandfather – Father – Son
    - Originally designed for tape backup, it works well for any hierarchical backup strategy
    - Define three sets of backups, such as daily, weekly and monthly
    - The daily, or son, backups are rotated on a daily basis
    - The weekly or father backups are rotated on a weekly
    - The monthly or grandphader are rotated on a montly

- ❑ The Tower of Hanoi
  - It is based on the mathematics of the Tower of Hanoi puzzle, with what is essentially a recursive method



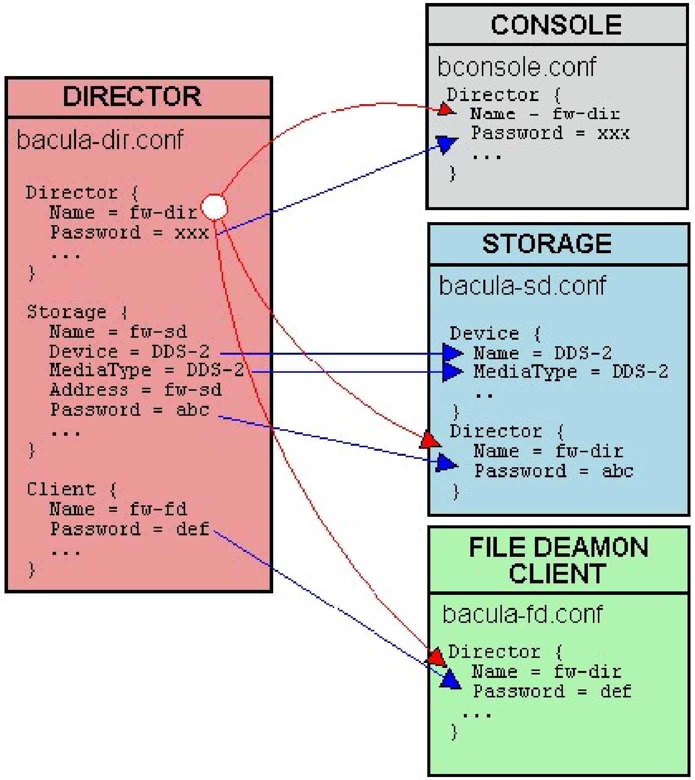
Tower of Hanoi Rotation Scheme

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Media Set	A		A		A		A		A		A		A		A	
		B				B				B				B		
				C								C				
								D								
																E

Return to Day 1

Bacula I

- Is a set of Open Source, computer programs
  - Manage backup
  - Recovery, and verification of computer data across a network of computers of different kinds
  - Bacula is by far the most popular Open Source program backup program
- Architectura
  - bacula-dir o bacula-director
    - This daemon co-ordinate all working of backup
  - bacula-sd
    - This daemon manage the information about the storage device that are availale to storage the backups
  - Bacula-fd
    - Run on machines or clients to be backuped
  - Console damon
    - A terminal (bconosle) or graphical (bat) daemon to control all works. These connect directly to bacula-dir daemon
  - Catalog database
    - Used for store all information related to the backup, including the file indexing
    - Mysql for large deployments, postgresql, mysql-lite (about 20 nodes)



DIRECTOR

bacula-dir.conf

```
Director {
  Name = fw-dir
  Password = xxx
  ...
}

Storage {
  Name = fw-sd
  Device = DDS-2
  MediaType = DDS-2
  Address = fw-sd
  Password = abc
  ...
}

Client {
  Name = fw-fd
  Password = def
  ...
}
```

CONSOLE

bconsole.conf

```
Director {
  Name - fw-dir
  Password = xxx
  ...
}
```

STORAGE

bacula-sd.conf

```
Device {
  Name = DDS-2
  MediaType = DDS-2
  ..
}
Director {
  Name = fw-dir
  Password = abc
}
```

FILE DEAMON CLIENT

bacula-fd.conf

```
Director {
  Name = fw-dir
  Password = def
 ...
}
```

Bacula II

- Install Bacula
  - On server you can compille directly form sources or the easiest way :

```
#apt-get install bacula-common bacula-console bacula-director-common bacula-fd bacula-sd libpq3 mysql-server mtx mt-st
#/etc/init.d/Bacula-director start
#/etc/init.d/Bacula-sd start
#/etc/init.d/Bacula-fd start
```

  - On Client Side

```
#apt-get install bacula-fd
#/etc/init.d/Bacula-fd start
```

  - Bacula use the 9102 and 9103 tcp ports

```
iptables -A OUTPUT -p tcp --dport 9102 -j ACCEPT
storage daemon
iptables -A OUTPUT -p tcp --dport 9103 -j ACCEPT
client daemon
```

Bacula III

- Configuration
  - The config file should be located at /etc/bacula/ directory
  - The Master config file for bacula director is bacula-director.conf
    - This File has a lot of directives (clients, schedulers, pools, storages)
    - Is a good practise to separate the configuration files and call

```
# Include below all yours jobs configuration files (remember add '@' at beginning)
 # Clients definition files
@/etc/bacula/conf.d/clients.conf
#Schedulers definition files
@/etc/bacula/conf.d/schedulers.conf
# Storages definition file
@/etc/bacula/conf.d/storages.conf
# Pools definition files
@/etc/bacula/conf.d/pools.conf
```

Bacula IV

- The client directive
  - defines and authorizes a customer to be backuped

```
Client {
  Name = wngw-fd
  Address = “machine_ip”
  FDPort = 9102
  Catalog = MyCatalog
  Password = “Very Long Password"
  File Retention = 90 days
  Job Retention = 6 months
  AutoPrune = yes
}
Client {
…..
}
```

- The Scheduler directive
  - Defines backup cycles

```
Schedule {
  Name = "WeeklyCycleSat"
  Run = Full 1st sat at 04:00
  Run = Differential 2nd-5th sat at 04:00
  Run = Incremental sun-fri at 04:00
}


# This schedule does the catalog. It starts after the
WeeklyCycle
Schedule {
  Name = "OneMonthFull"
  Run = Full 1st sun at 23:05
}

Schedule {
  Name = "WeeklyCycleAfterBackup"
  Run = Full sun-sat at 23:10
}
```

Bacula V

- The storages directive
  - Defines and authorizes the access to bacula-sd daemon

```
Storage {
  Name = File
# Do not use "localhost" here
  Address = 10.10.0.22
  SDPort = 9103
  Password = " very _long _passwd"
  Device = FileStorage
  Media Type = File
}

Storage {
  Name = TSM3500-LTO3
  Address = 10.10.0.22
  SDPort = 9103
  Password = “very _long _passwd"
  Device = ULT3580-TD3
  Media Type = LTO-3
  Autochanger = yes
  Maximum Concurrent Jobs = 1
}
```

- The pools directive
  - Defines Logical storage areas to makes backups

```
Pool {
       Name = Full
       Label Format = "BckF"
       Pool Type = Backup
       Recycle = yes AutoPrune = yes          # Prune
expired volumes
       Volume Retention = 6 months
       Maximum Volume Jobs = 50      #Similar to clients
number
       Maximum Volumes = 300
}
Pool {
       Name = Diff
       Label Format = "BackD"
       Pool Type = Backup
       Recycle = yes
       AutoPrune = yes
       Volume Retention =  30 days
       Maximum Volume Jobs = 50
       Maximum Volumes = 500
}
```

Bacula VI

- The Filesets directive
  - Defines the directories7files to be backuped or excluded, and order options

```
FileSet {
        Name = "FullServices"
        Include {
                Options {
                        signature = MD5
                        compression=GZIP
                }
                File = /
        }

        Exclude {
                File = /gpfs
                File = /mnt
                File = /proc
                File = /sys
                File = /tmp
                File = /.journal
                File = /.fsck
                File = /nfs4
                File = /swapfile
        }
}
```

- The Job directive
  - Merge previous directives to spaln the job

```
JobDefs {
         Name = "GridFtp"
         Enabled = yes
         FileSet = "ExcludeGPFS"
         Schedule = "OneMonthFull"
         Write Bootstrap = "/var/lib/bacula/gridftp.bsr"
         Storage = TSM3500-LTO3
         Type = Backup
         Level = Full
         Pool = Full
         Priority = 10
         Messages = Standard
         Reschedule On Error = yes
         Reschedule Interval = 1 hour
         Reschedule Times = 1
}

Job {
    Name = "Backpool03"
    JobDefs = "GridFtp"
    Client = pool03-fd
}
```

Bacula VII

- The Master config file for bacula storage is bacula-sd.conf
  - This makes reference to physical storage device

```
Device {
  # The TSM3100's second tape drive
  Name = ULT3580-TD5
  Archive Device = /dev/nst1
  Device Type = Tape
  Media Type = LTO-5
  Autochanger = Yes
  # Changer Device = <inherited from Changer>
  Alert Command = "sh -c '/usr/sbin/tapeinfo -f /dev/sg6 | /bin/sed
-n /TapeAlert/p"
  Drive Index = 0
  RemovableMedia = yes
  Random Access = no
  Maximum Spool Size = 80gb
  Maximum Job Spool Size = 40gb
  Spool Directory = /backup/spool
  AutomaticMount = Yes;
}
```

```
Device {
  Name = FileStorage
  Media Type = File
  Archive Device = /Bacula/Default
  LabelMedia = yes;                   # lets Bacula label
unlabeled media
  Random Access = Yes;
  AutomaticMount = yes;               # when device
opened, read it
  RemovableMedia = no;
  AlwaysOpen = no;
}

Autochanger {
  Name = TSM3500
  Device = ULT3580-TD3
  Device = ULT3580-TD5
  Changer Device = /dev/sg7
  Changer Command = "/etc/bacula/scripts/mtx-
changer %c %o %S %a %d"
                    # %c = changer device
                    # %o = command
(unload|load|loaded|list|slots)
                    # %S = slot index (1-based)
                    # %a = archive device (i.e., /dev/sd* name
for tape drive)

                    # %d = drive index (0-based)
}
```

Bacula VIII

- The Master config file for bacula storage is bacula-fd.conf
  - This makes reference to physical storage device

```
FileDaemon {                          # this is me
  Name = bacula-fd
  FDport = 9102                  # where we listen for the
director
  WorkingDirectory = /var/lib/bacula
  Pid Directory = /var/run/bacula
  Maximum Concurrent Jobs = 20
  FDAddress = 10.10.0.22
  #FDAddress = 127.0.0.1
}

# Send all messages except skipped files back to
Director
Messages {
  Name = Standard
  director = bacula-dir = all, !skipped, !restored
}
```

Bacula IX

- Example Create a label/Volume on tapes
  - To list the available tapes

```
#/etc/bacula/scripts/mtx-changer /dev/sg7 listall 0 /dev/nst1 0 |more
```

  - Inicialize the tapes

```
#/etc/bacula/scripts/mtx-changer /dev/sg7 load $i /dev/$tape $drive
#mt -f /dev/nst0 rewind
#mt -f /dev/nst0 weof
#/etc/bacula/scripts/mtx-changer /dev/sg7 unload $i /dev/$tape $drive##
```

  - Select the storage resource

```
The defined Storage resources are:
     1: File
     2: TSM3500-LTO3
     3: TSM3500-LTO5
Select Storage resource (1-3): 3
```

  - Select the drive and volume name and slot

```
Enter autochanger drive[0]:
Enter new Volume name: B00680LV
Enter slot (0 or Enter for none): 28
```

  - Select the pool

```
Select the Pool (1-8): 7
Sending label command for Volume "B00680LV" Slot 28 ...
3000 OK label. VolBytes=64512 DVD=0 Volume="B00680LV" Device="ULT3580-TD5" (/dev/nst1)
Catalog record for Volume "B00680LV", Slot 28  successfully created.
```

Bacula X

- Example add machine to be backuped
  - Install bacula, and bacula-client packages

```
rpm - Uvh bacula-5.0.3-1.el5.pp.x86_64 bacula-client-5.0.3-
1.el5.pp.x86_64 (Redhat)
apt-get install bacula-client  (Debian, Ubuntu)
```

  - Set the correct values on bacula-fd.conf file and start the service

```
#/etc/init.d/bacula-fd start
```

  - Add the client to bacula director (clients.conf)

```
Client {
  Name = machine-fd
  Address = machine_ip
  FDPort = 9102
  Catalog = MyCatalog
  Password = “connection_passwd"
  File Retention = 90 days
  Job Retention = 6 months
  AutoPrune = yes
}
```

Bacula XI

- Add to bacula jobs, selecting the de desirable choices (scheduling, pools, priority…)

```
JobDefs {
        Name = "Services"
        Enabled = yes
        FileSet = "FullServices"
        Schedule = "WeeklyCycleSun"
        Write Bootstrap = "/var/lib/bacula/services.bsr"
        Full Backup Pool = Full
        Incremental Backup Pool = Incr
        Differential Backup Pool = Diff
        Storage = TSM3500-LTO3
        Type = Backup
        Level = Full
        Pool = Full
        Priority = 10
        Messages = Standard
        Reschedule On Error = yes
        Reschedule Interval = 1 hour
        Reschedule Times = 1
}
Job {
    Name = "BackCerbero"
    JobDefs = "Services"
    Client = cerbero-fd
}
```

- Restart de director service

```
#/etc/init.d/bacula-fdirector restart
```