Introduction to Linux for Bioinformatics

Managing data

Joachim Jacob
5 and 12 May 2014

Bioinformatics data

Historically, bioinformatics has always used text files to store data.

Homo sapiens CD99 molecule, mRNA (cDNA clone MGC:19734 IMAGE:3606974), complete cds

GenBank: BC010109.2

FASTA Graphics

Go to:

```
LOCUS       BC010109                1255 bp    mRNA    linear   PRI 24-JUL-2006
DEFINITION  Homo sapiens CD99 molecule, mRNA (cDNA clone MGC:19734
            IMAGE:3606974), complete cds.
ACCESSION   BC010109
VERSION     BC010109.2  GI:33991438
KEYWORDS    MGC.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1255)
  AUTHORS   Strausberg,
            Klausner,R
            Altschul,S
            Hopkins,R.
            Diatchenko,
            Stapleton,
            Scheetz,T.
            Carninci,P
            Abramson,R
            McKernan,K
```

Genbank record

PDB file excerpt

4B4T.pdb

```
CISPEP   8 GLY U   96    PRO U   97          0        12.74
CISPEP   9 LYS V  148    GLY V  149          0        10.78
CISPEP  10 PRO W   22    ARG W   23          0        -8.01
CRYST1    1.000    1.000    1.000  90.00  90.00  90.00 P 1
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      1.000000  0.000000  0.000000        0.00000
SCALE2      0.000000  1.000000  0.000000        0.00000
SCALE3      0.000000  0.000000  1.000000        0.00000
ATOM      1  N   LEU 1  -9     291.873 272.970 292.148  1.00  0.
ATOM      2  CA  LEU 1  -9     292.720 272.665 293.227  1.00  0.
ATOM      3  C   LEU 1  -9     293.269 271.278 292.826  1.00  0.
ATOM      4  O   LEU 1  -9     294.327 271.206 292.179  1.00  0.
ATOM      5  CB  LEU 1  -9     291.941 272.689 294.543  1.00  0.
ATOM      6  CG  LEU 1  -9     291.406 274.052 294.990  1.00  0.
ATOM      7  CD1 LEU 1  -9     290.527 273.906 296.222  1.00  0.
ATOM      8  CD2 LEU 1  -9     292.551 275.013 295.268  1.00  0.
ATOM      9  N   LYS 1  -8     292.589 270.185 293.321  1.00  0.
ATOM     10  CA  LYS 1  -8     292.957 268.825 293.012  1.00  0.
ATOM     11  C   LYS 1  -8     291.730 268.185 292.594  1.00  0.
ATOM     12  O   LYS 1  -8     290.627 268.758 292.701  1.00  0.
ATOM     13  CB  LYS 1  -8     293.602 268.155 294.227  1.00  0.
ATOM     14  CG  LYS 1  -8     294.880 268.827 294.700  1.00  0.
ATOM     15  CD  LYS 1  -8     296.015 268.615 293.711  1.00  0.
ATOM     16  CE  LYS 1  -8     297.308 269.238 294.211  1.00  0.
ATOM     17  NZ  LYS 1  -8     298.397 269.143 293.201  1.00  0.
```

```
HMMER3/b [3.0 | March 2010]
NAME  Histones model
LENG  251
ALPH  amino
RF    no
CS    no
MAP   yes
DATE  Tue Feb 28 14:11:20 2012
NSEQ  6
EFFN  1.004883
CKSUM 427625644
STATS LOCAL MSV      -11.2973  0.70287
STATS LOCAL VITERBI  -12.0628  0.70287
STATS LOCAL FORWARD   -4.6693  0.70287
HMM          A        C        D        E        F        G        H        I        K
S        T        V        W        Y
            m->m     m->i     m->d     i->m     i->i     d->m     d->d
COMPO   2.11140  4.62155  3.14861  2.71408  3.75567  3.10072  3.81893  3.14056  2.06706
2.58900  2.75668  2.77637  5.30704  3.87841
         2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741
2.37887  2.77519  2.98518  4.58477  3.61503
         0.02270  4.19253  4.91488  0.61958  0.77255  0.00000        *
      1  3.25495  4.75900  4.41277  4.04275  3.39324  4.08493  4.70396  2.60315  3.79035
3.63187  3.57200  2.64439  5.26941  4.05464      1 - -
         2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741
2.37887  2.77519  2.98518  4.58477  3.61503
         0.14016  4.19253  2.15703  0.61958  0.77255  0.48576  0.95510
      2  1.03405  4.18028  3.55056  3.29115  4.31149  2.97979  4.32632  3.64354  3.33291  3.41228  4.26990  3.39467  3.74311  3.62328  3.61504
1.93679  2.75614  3.13074  5.68663  4.47676      2 - -
         2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801
2.37887  2.77519  2.98518  4.58477  3.61503
         0.12270  4.07787  2.31744  0.61958  0.77255  0.41345  1.08282
```

HMM profile

JAKONIX

VIB

NGS data

The NGS machines produce a lot of data, stored in **plain text** files. These files are multiple gigabytes in size.

Tips for managing (NGS) data

1. *When you move the data, do it in its smallest form.*

→ **Compress** the data.

2. *When you decompress the data, leave it where it is.*

→ **Symbolic links** point to the data in different folders.

3. Provide enough storage for your data.

→ choose your **file system type** wisely

Compression: tools in Linux

Compression Tools	
Lrzip	Achieve very high compression ratios and speed when used with large files
lbzip2	Multi-threaded implementation of bzip2, suited for serial and parallel processing
7-Zip	File archiver with a high compression ratio
XZ Utils	Successor to the Lempel-Ziv/Markov-chain Algorithm compression format
bzip2	Freely available, patent free, high-quality data compressor
gzip	Provides the standard GNU file compression utilities
PeaZip	Cross-platform portable file archiver

But some more exist (e.g. freearc, dtrx). **gzip** and **bzip2** are the most used and fairly performant.

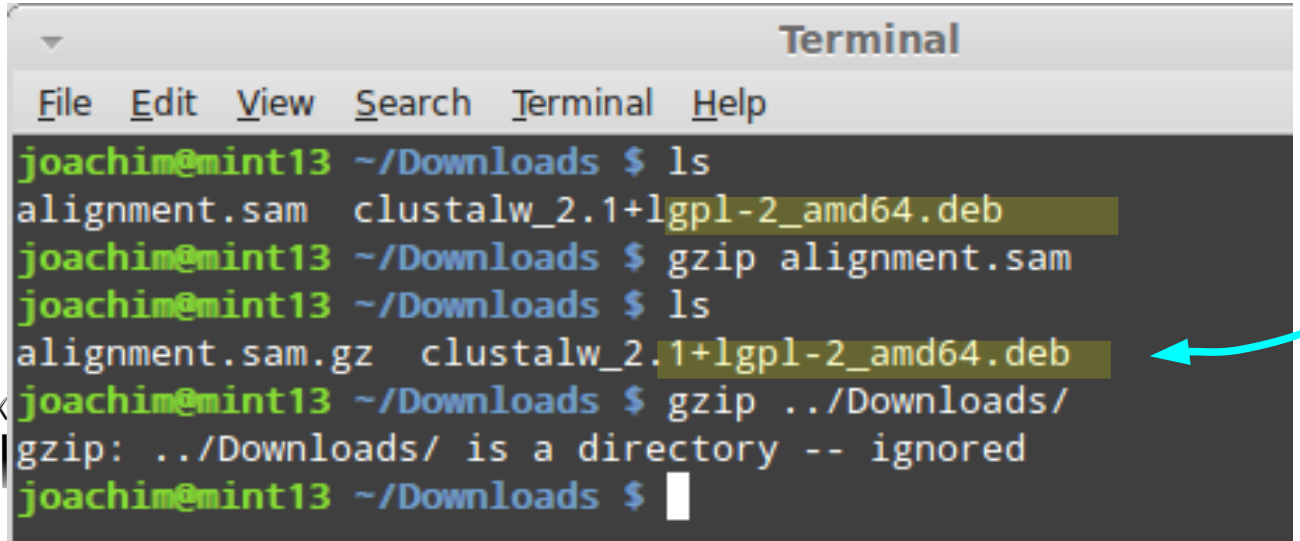
http://www.linuxlinks.com/article/20110220091109939/CompressionTools.html

Tips

Widely used compression tools:

- GNU zip (`gzip`)
- Block Sorting compression (`bzip2`)

Typically, compression tools work on one file. How to compress complete directories and their contents?



Tar without compression

Tar (Tape Archive) is a tool for **bundling a set of files or directories into a single archive.** The resulting file is called a tar ball.

Syntax to create a tarball:

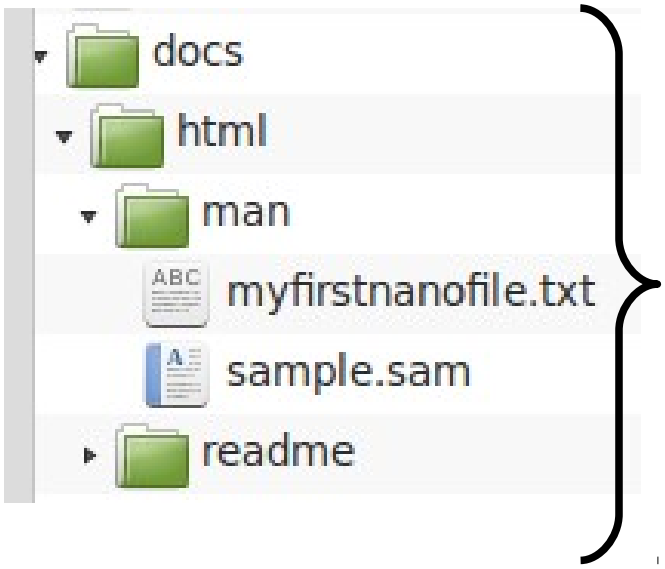
```
$ tar -cf archive.tar file1 file2
```

Syntax to extract:

```
$ tar -xvf /path/to/archive.tar
```

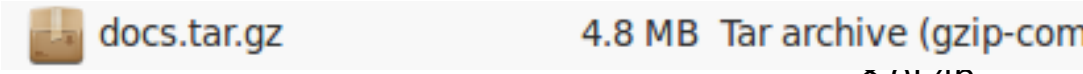
Compression: a typical case

Archiving and compression mostly occur together. The most used formats are **tar.gz** or **tar.bz.** These files are the result of **two** processes.

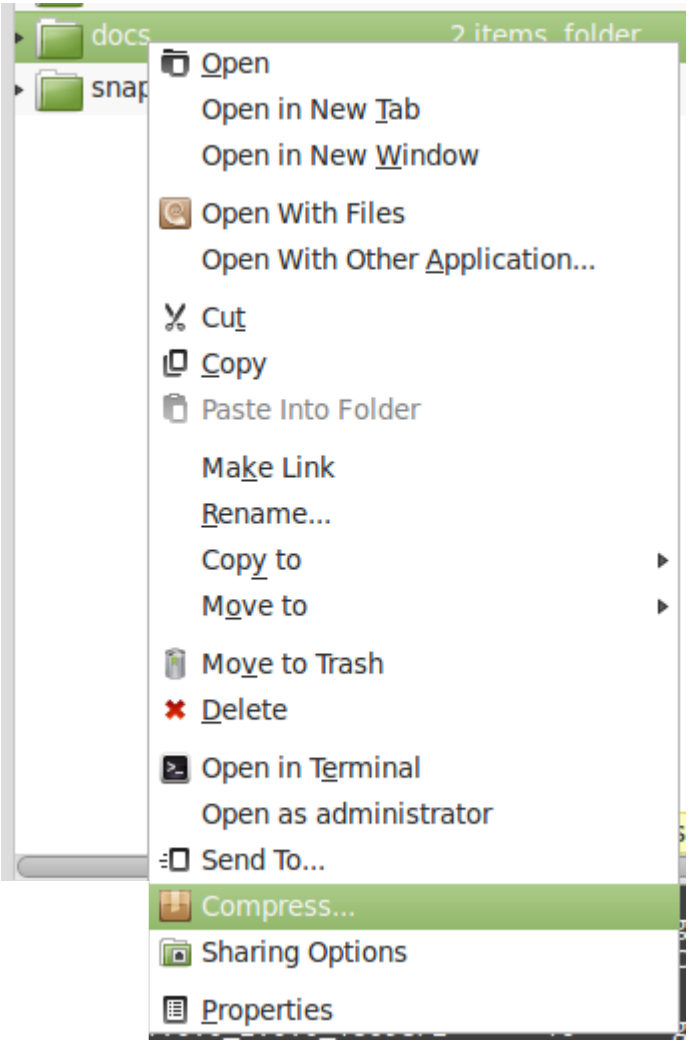


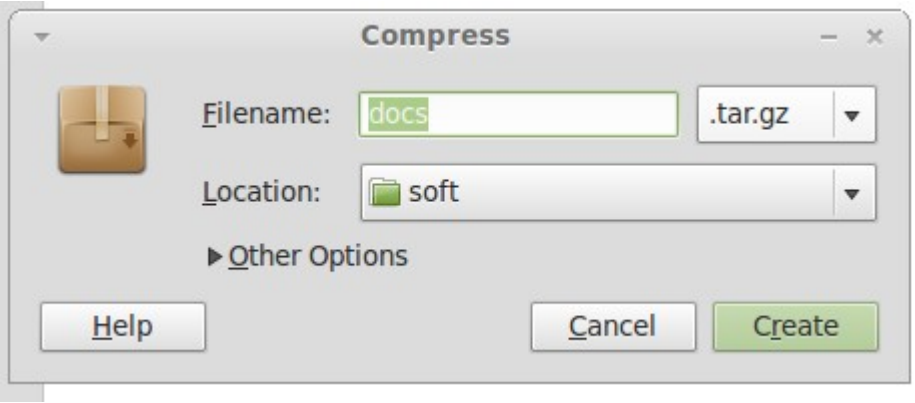
1. Archiving (tar)

2. Compressing (gzip or bzip2)

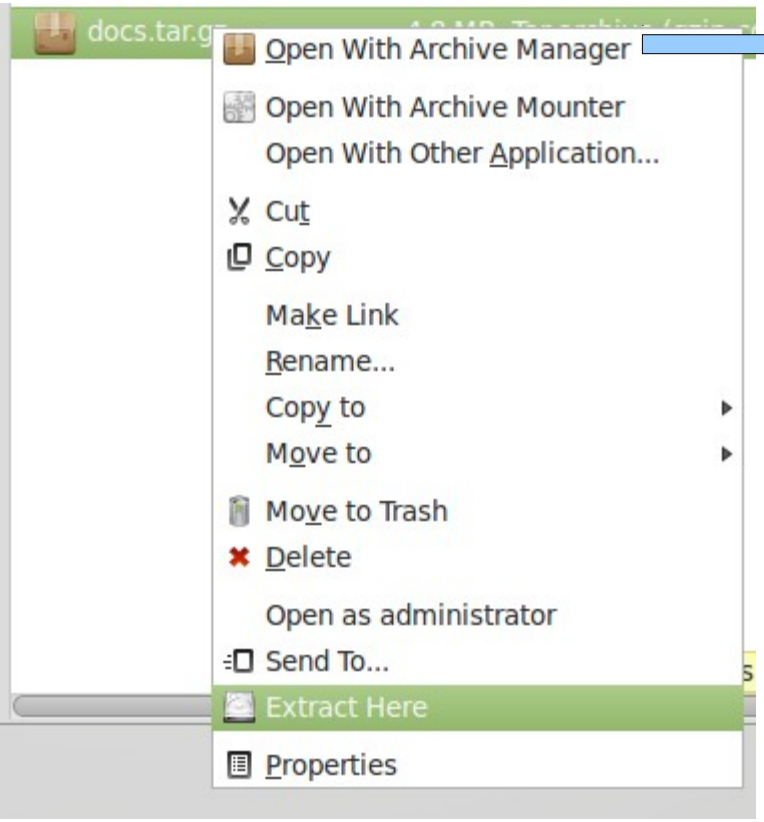


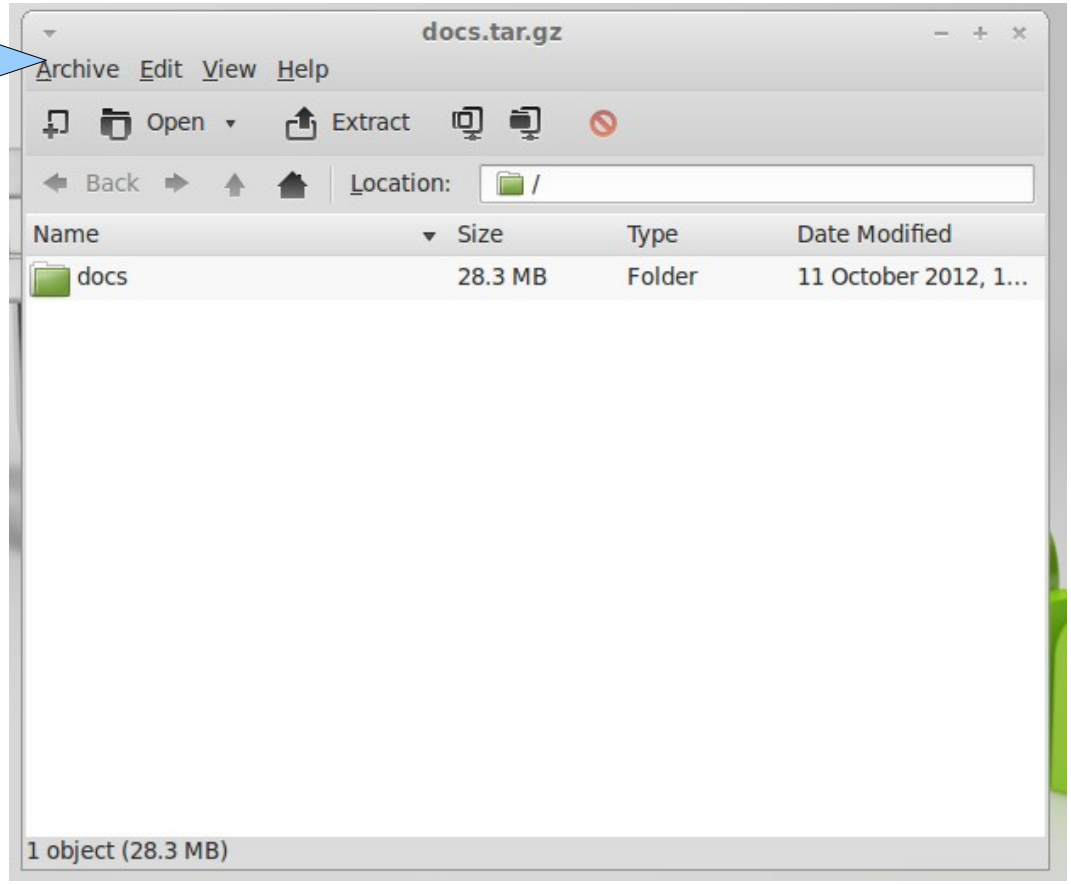
Compression: on your desktop





Compression: on your desktop





Compression: on the command line

Tar is the tool for creating .tar archives, but it can compress in one go, with the z or j option.

Creating a compressed tar archive:

```
$ tar cvfz  mytararchive.tar.gz  docs/
$ tar cvfj  mytararchive.tar.bz  docs/
```

create

Compression technique

Decompressing a compressed tar archive

```
$ tar xvfz mytararchive.tar.gz
$ tar xvfj mytararchive.tar.bz
```

extract

files

verbose

De-/compression

To compress one or more files:

```
$ gzip [options] file
$ bzip2 [options] file
```

To decompress one or more files:

```
$ gunzip [options] file(s)
$ bunzip2 [options] file(s)
```

Every file will be compressed, and tar.gz or tar.bz appended to it.

Tips

1. Do you have to uncompress a big text file to read it? No! Some tools allow to **read compressed files** (instead of first unpacking then reading). Time saver!

```
$ zcat file(s)
$ bzcat file(s)
```

2. Compression is always a **balance** between time and compression ratio. Gzip is faster, bzip2 compresses harder.

If compression is important to you: benchmark!

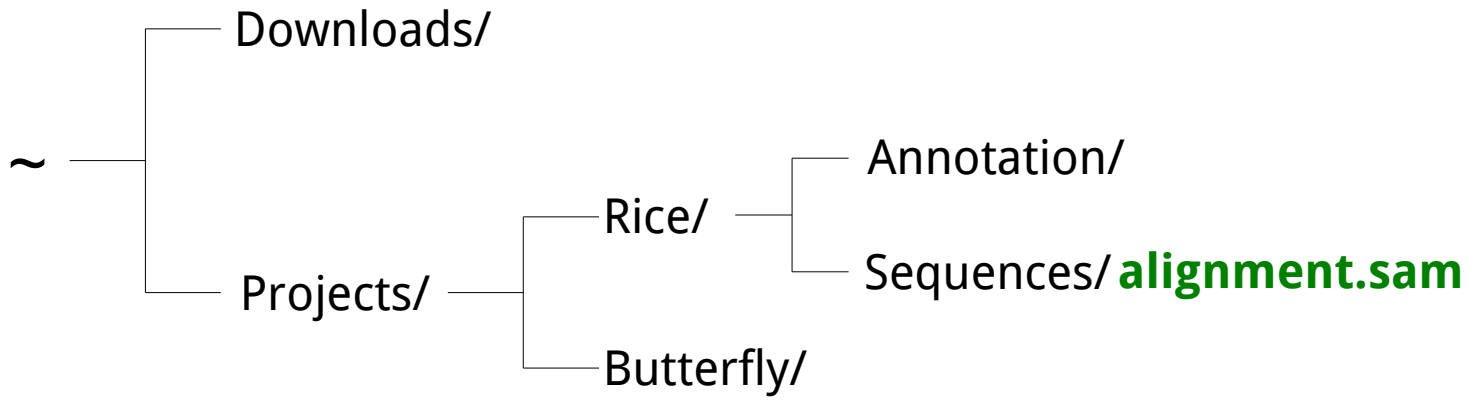
Exercise

→ a little compression exercise.

Symlinks

Something very convenient!

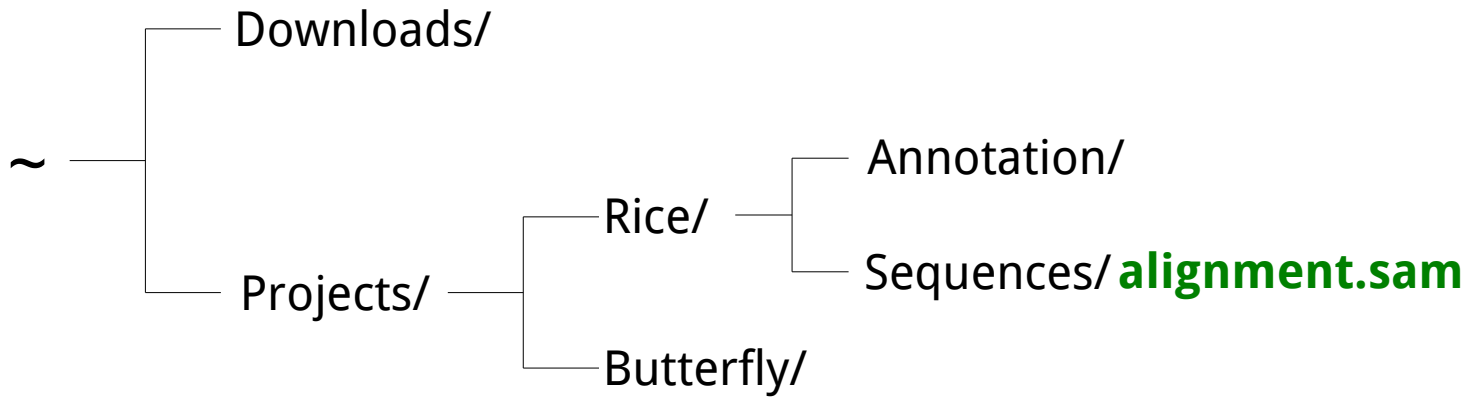
A **symbolic link** (or symlink) is a *file* which points to the location of another file. You can do anything with the symlink that *you can do on the original file*. But when you move the original file from its location, the symlink is 'dead'.



Symlinks

To create a symlink, move to the folder in where the symlink must be created, and execute `ln`.

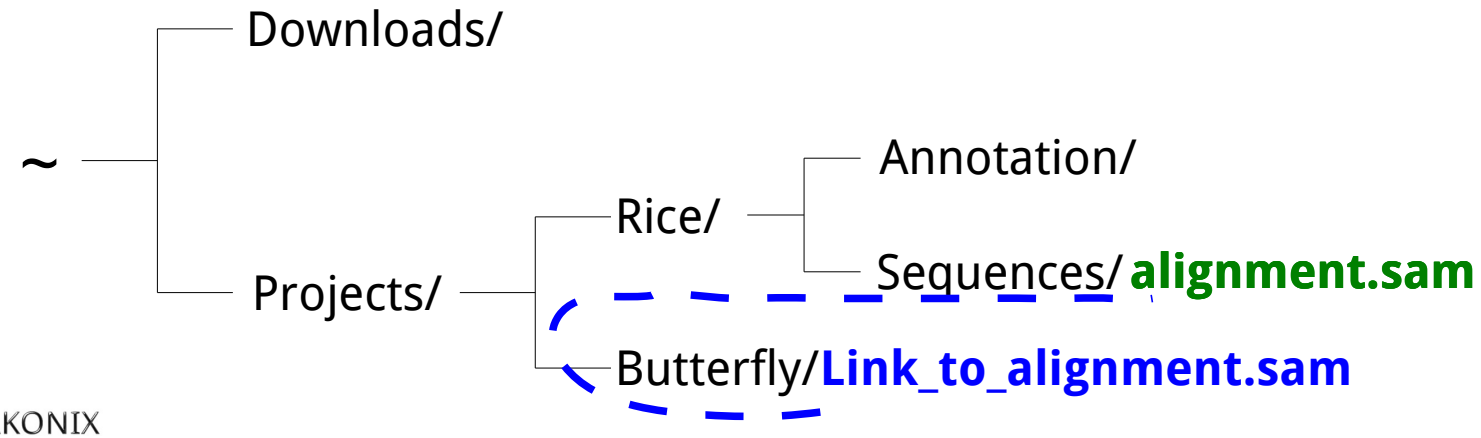
```
~/Projects $ cd Butterfly
~/Butterfly $ ln -s ../Rice/Sequences/alignment.sam
Link_to_alignment.sam
```



Symlinks

The symlink is created. You can check with `ls`.
To delete a symlink, use `unlink`.

```
~/Projects $ cd Butterfly
~/Butterfly $ ln -s ../Rice/Sequences/alignment.sam
Link_to_alignment.sam
~/Butterfly $ ls -lh Link_to_alignment.sam
lrwxrwxrwx 1 joachim joachim 44 Oct 22 14:47
Link_to_alignment.sam -> ../Sequences/alignment.sam
```



Exercise

→ a little symlink exercise

Disks and storage

If you dive into bioinformatics, you will have to manage disks and storage.

Two types of disks

- solid state disks

Low capacity, high speed, random writes

- spinning hard disks

High capacity, 'normal' speed, sequential writes.

http://en.wikipedia.org/wiki/Solid-state_drive
http://en.wikipedia.org/wiki/Hard_disk

A disk is a device

Via the terminal, show the disks using

```
$ sudo fdisk -l
[sudo] password for joachim:

Disk /dev/sda: 13.4 GB, 13408141312
bytes
...
Disk /dev/sdb: 3997 MB, 3997171712 bytes
...
```

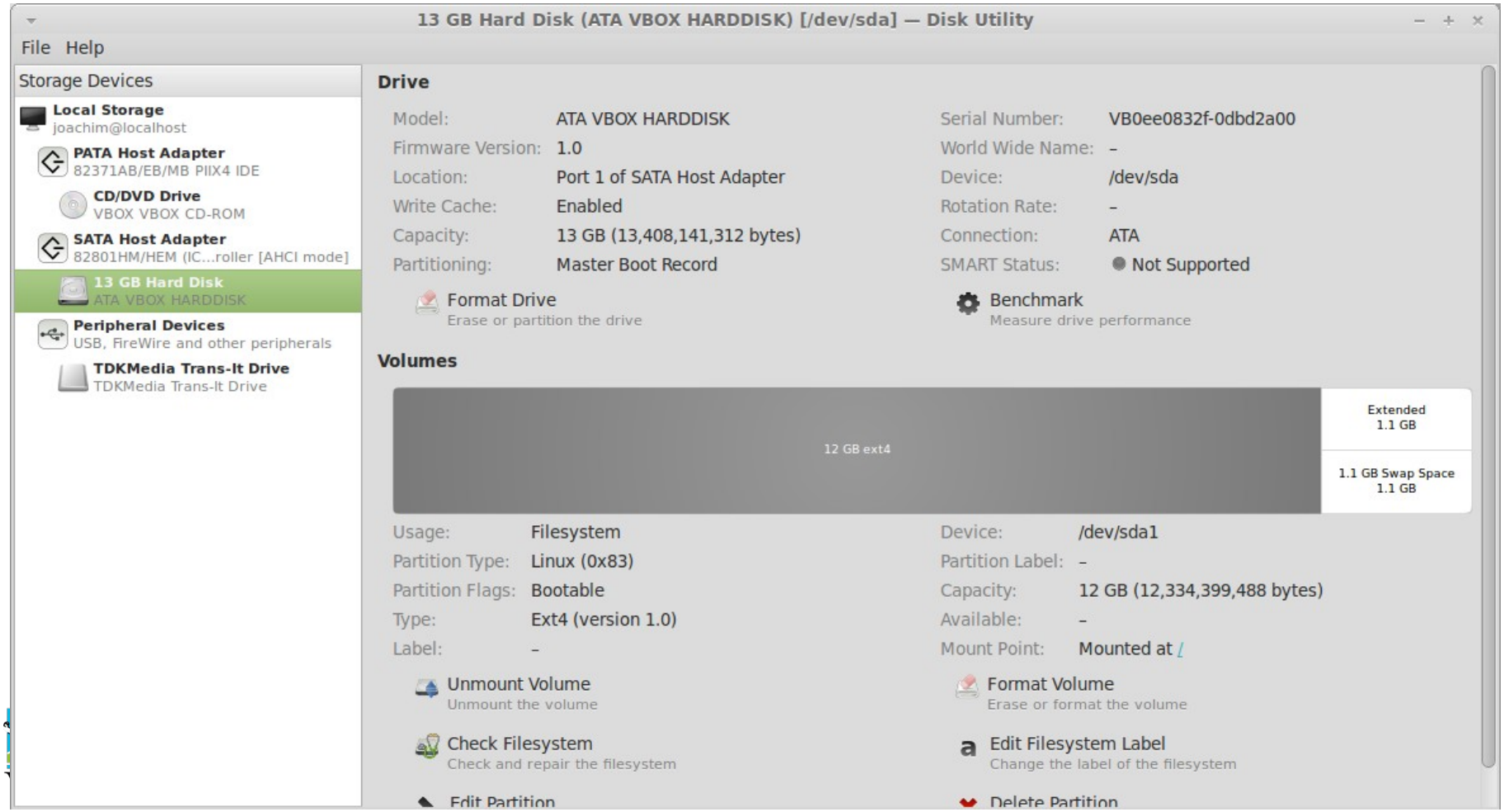
A disk is divided into partitions

A disk can be divided in parts, called partitions.

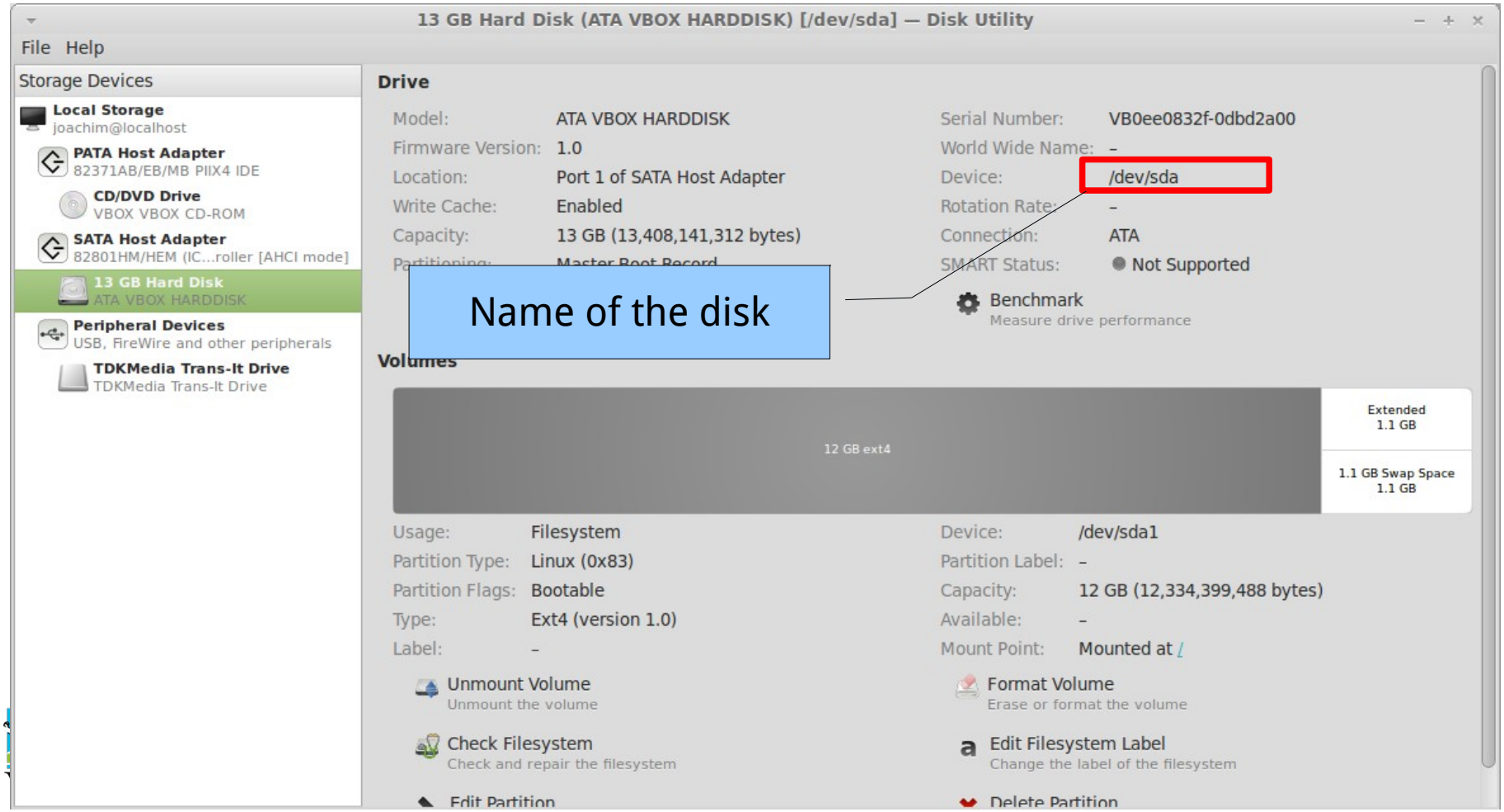
An **internal disk** which runs an operating system is usually divided in partitions, one for each functions.

An **external disk** is usually not divided in partitions. (but it can be partioned).

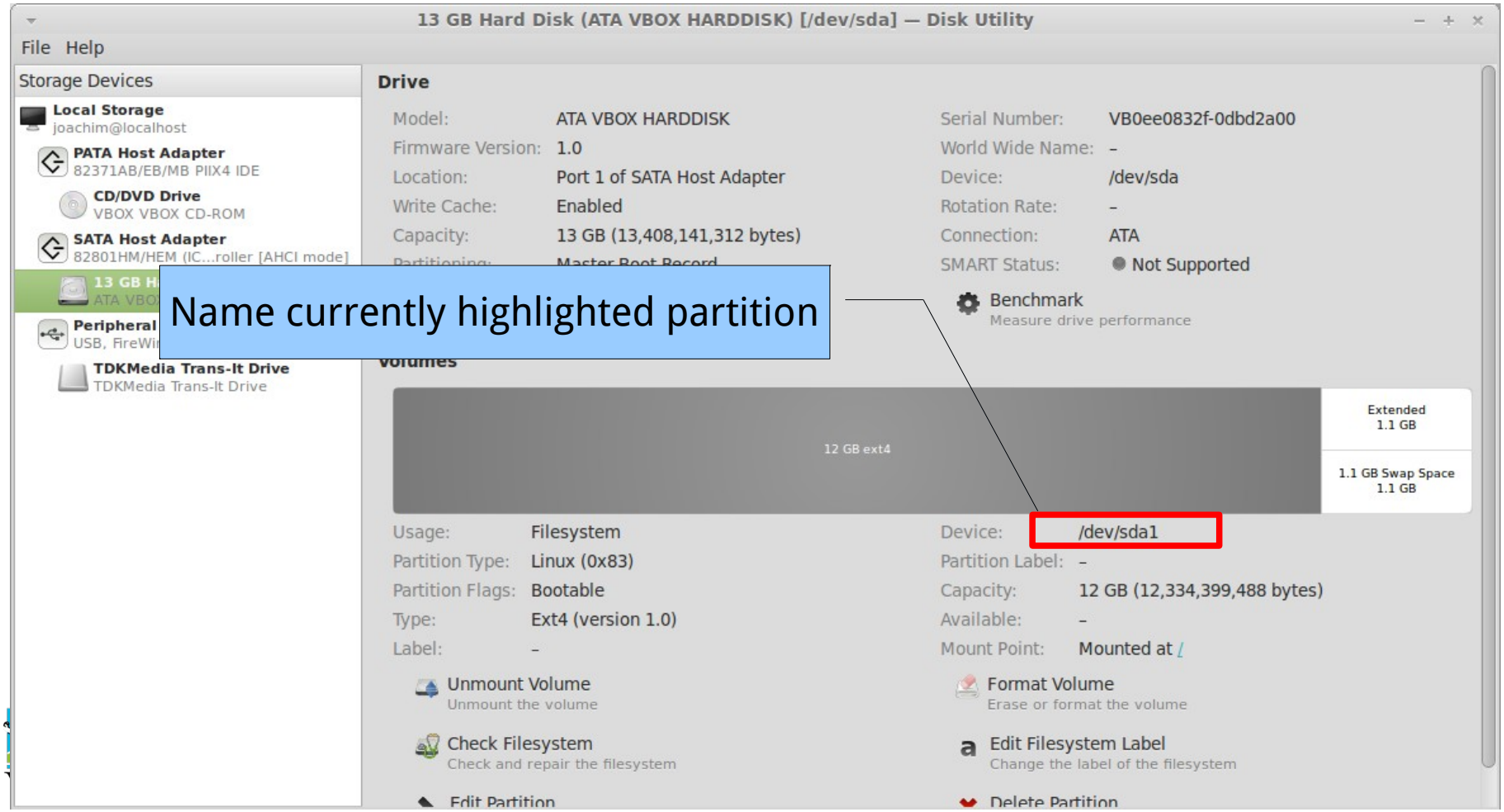
Check out the disk utility tool



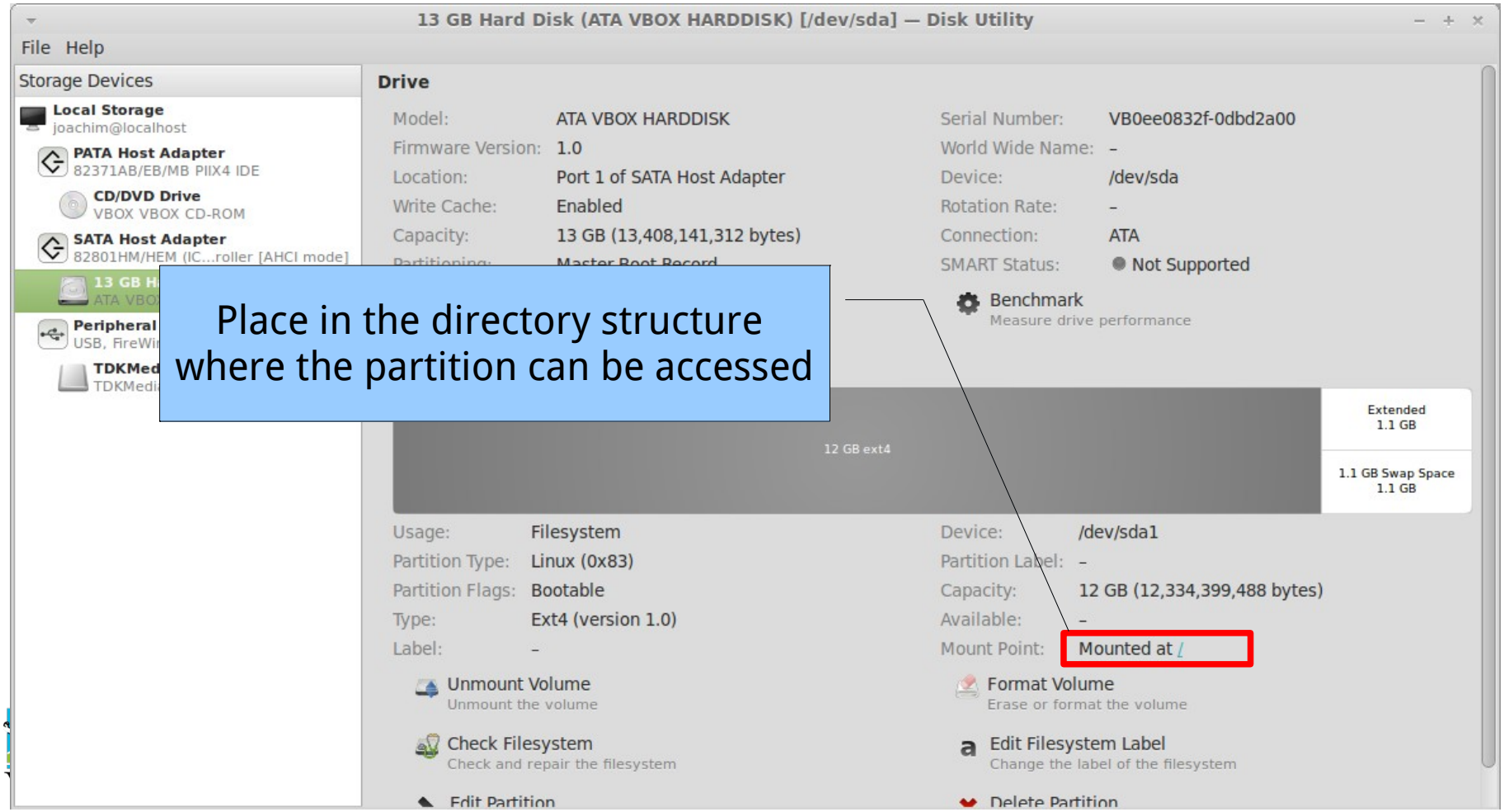
The system disk



The system disk

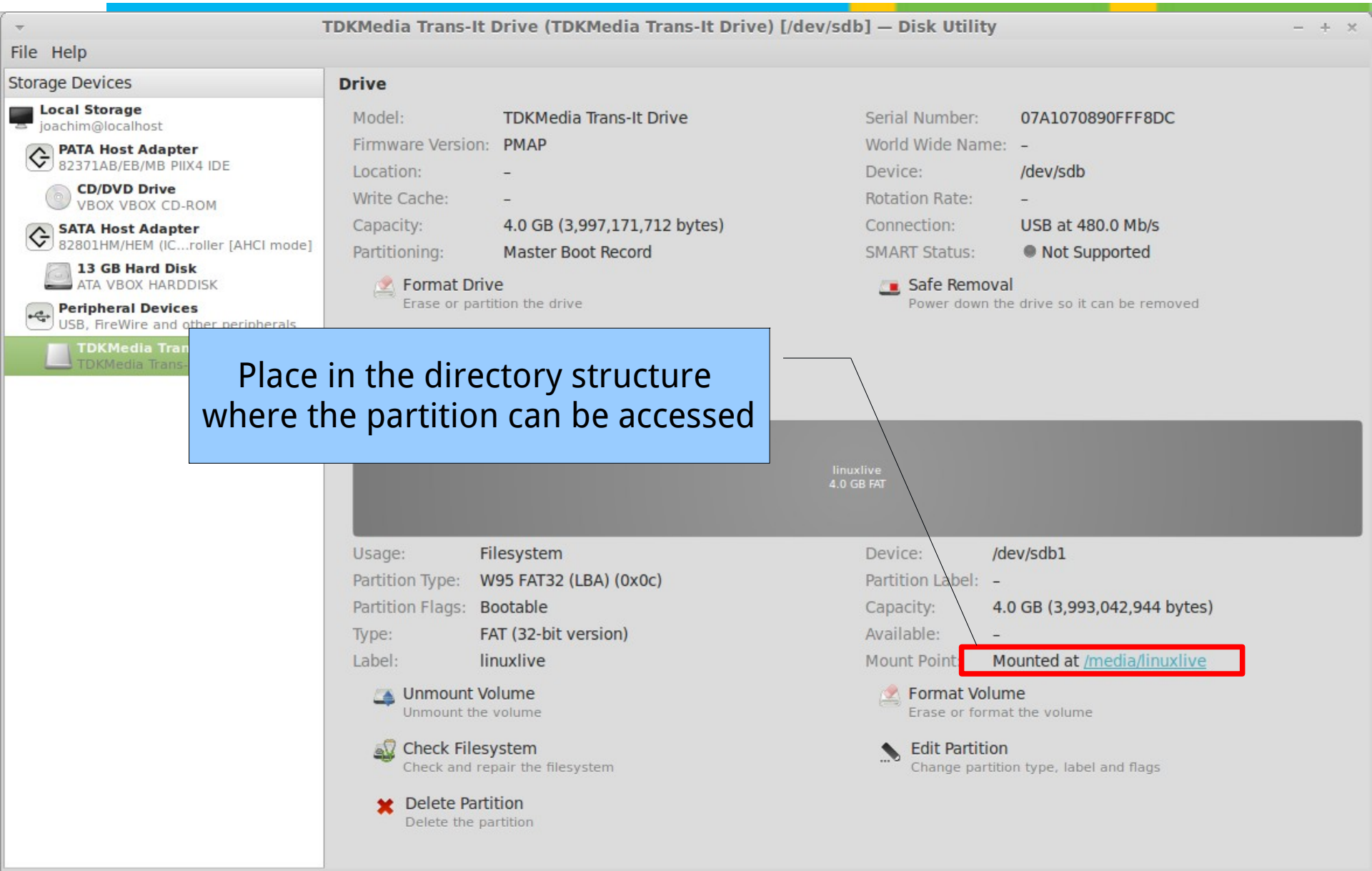


The system disk



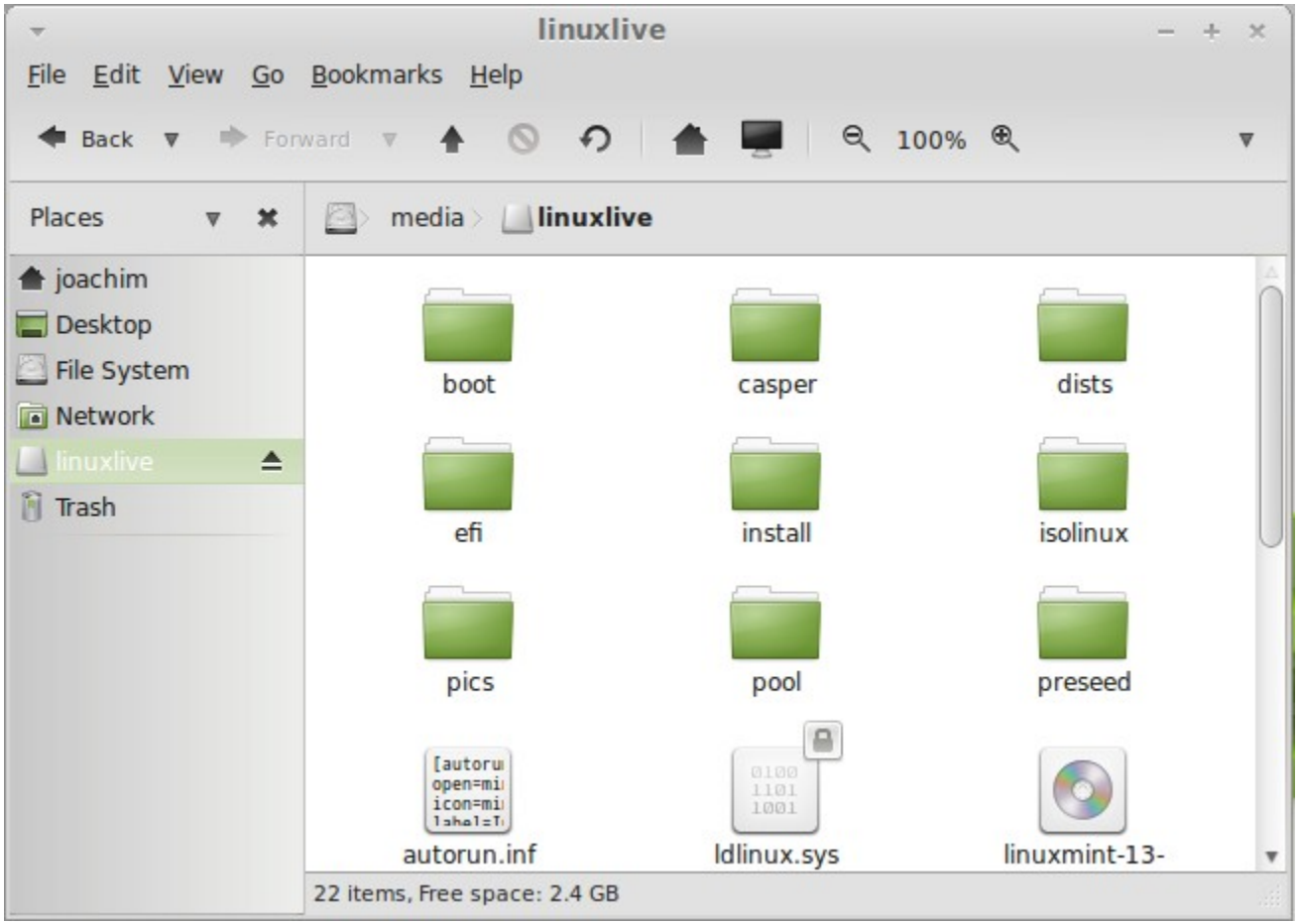
Place in the directory structure where the partition can be accessed

An example of an USB disk



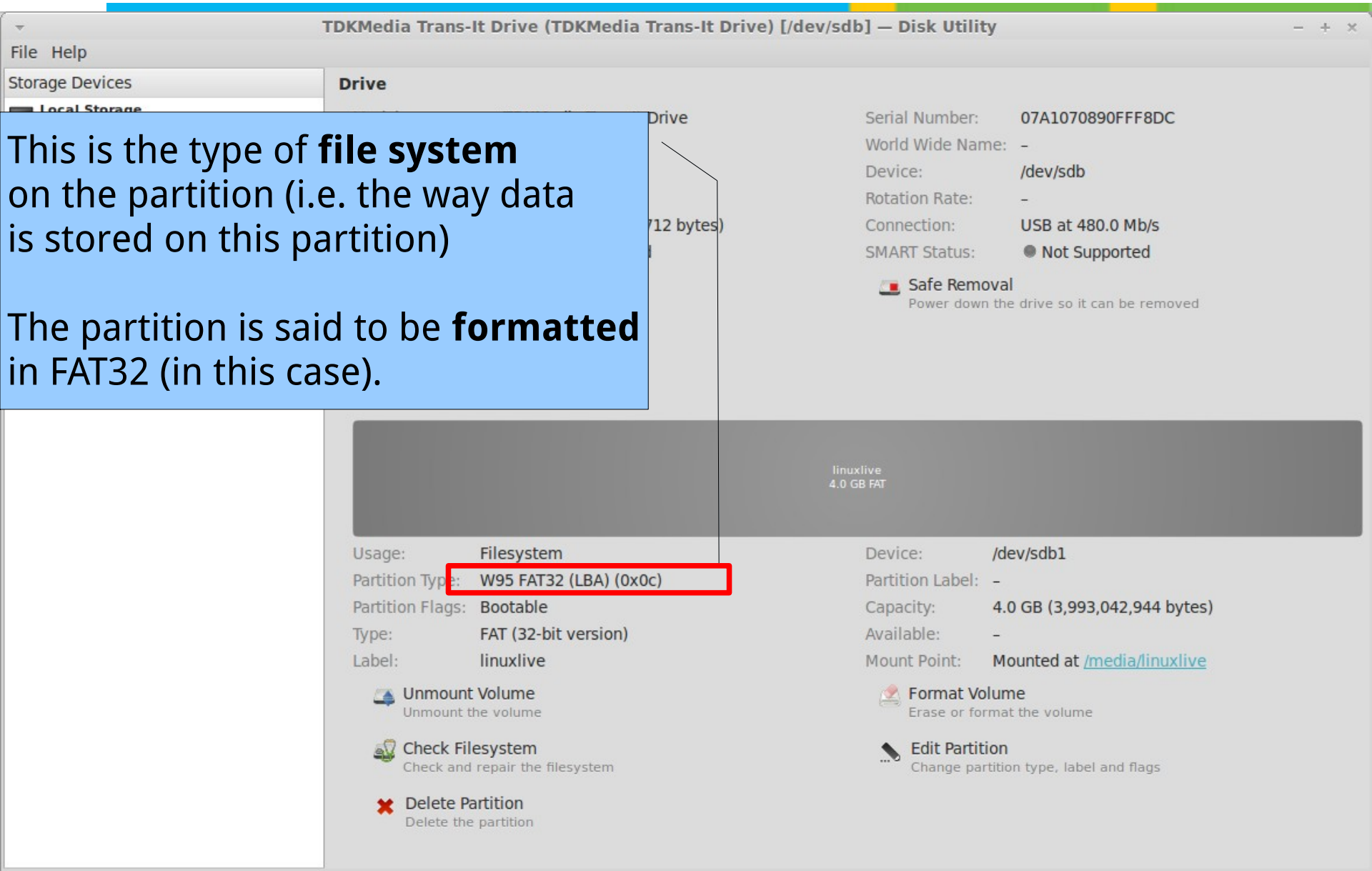
An example of an USB disk

The USB disk is 'mounted' automatically on the directory tree under **/media**.



JAKONIX
VIB

An example of an USB disk



This is the type of **file system** on the partition (i.e. the way data is stored on this partition)

The partition is said to be **formatted** in FAT32 (in this case).

File system formats

By default, many USB flash disks are formatted in **FAT32**.

Other types are NTFS, ext4, ZFS.

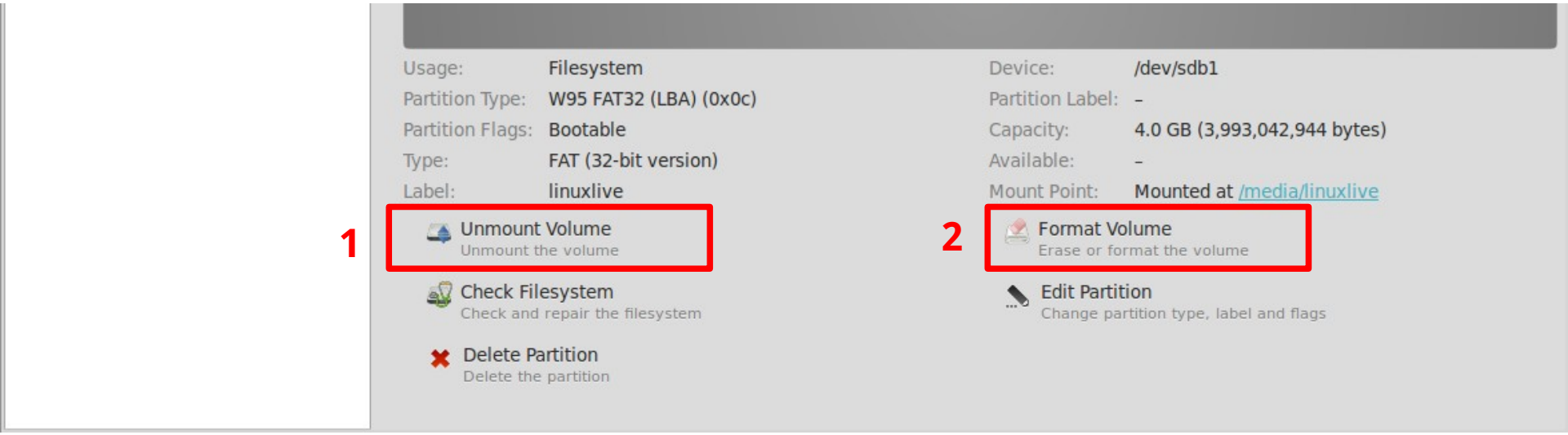
FAT32 – max 4GB files
NTFS – maximum portability (also for use under windows)
Ext4 – default file system in Linux,
Btrfs – the next default file system in Linux in the near future.

http://en.wikipedia.org/wiki/File_system#File_systems_and_operating_systems

Example: formatting a USB disk

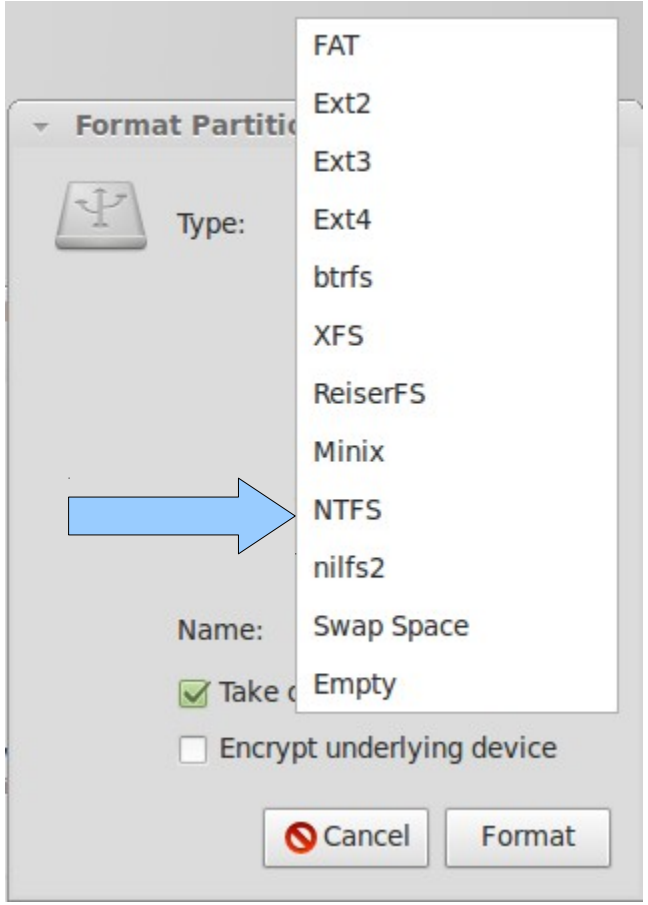
First **unmount** the device (red box 1 below). Now the USB disk can not be accessed anymore.

Next, choose **format** the device (red box 2 below).



Format disks with disk utility

Choose the type of file system you want to be on that device.



Format disks with disk utility

TDKMedia Trans-It Drive (TDKMedia Trans-It Drive) [/dev/sdb] — Disk Utility

File Help

Storage Devices

- **Local Storage** joachim@localhost
- **PATA Host Adapter** 82371AB/EB/MB PIIX4 IDE
  - **CD/DVD Drive** VBOX VBOX CD-ROM
- **SATA Host Adapter** 82801HM/HEM (IC...roller [AHCI mode]
  - **13 GB Hard Disk** ATA VBOX HARDDISK
- **Peripheral Devices** USB, FireWire and other peripherals
  - **TDKMedia Trans-It Drive** TDKMedia Trans-It Drive

Drive

Model:	TDKMedia Trans-It Drive	Serial Number:	07A1070890FFF8DC
Firmware Version:	PMAP	World Wide Name:	–
Location:	–	Device:	/dev/sdb
Write Cache:	–	Rotation Rate:	–
Capacity:	4.0 GB (3,997,171,712 bytes)	Connection:	USB at 480.0 Mb/s
Partitioning:	Master Boot Record	SMART Status:	Not Supported

Format Drive — Erase or partition the drive

Safe Removal — Power down the drive so it can be removed

Benchmark — Measure drive performance

Volumes

test
4.0 GB NTFS

Usage:	Filesystem	Device:	/dev/sdb1
Partition Type:	W95 FAT32 (LBA) (0x0c)	Partition Label:	–
Partition Flags:	Bootable	Capacity:	4.0 GB (3,993,042,944 bytes)
Type:	NTFS	Available:	–
Label:	test	Mount Point:	Mounted at /media/test

Unmount Volume — Unmount the volume

Format Volume — Erase or format the volume

Check Filesystem — Check and repair the filesystem

Edit Partition — Change partition type, label and flags

Delete Partition — Delete the partition

Format disks with disk utility

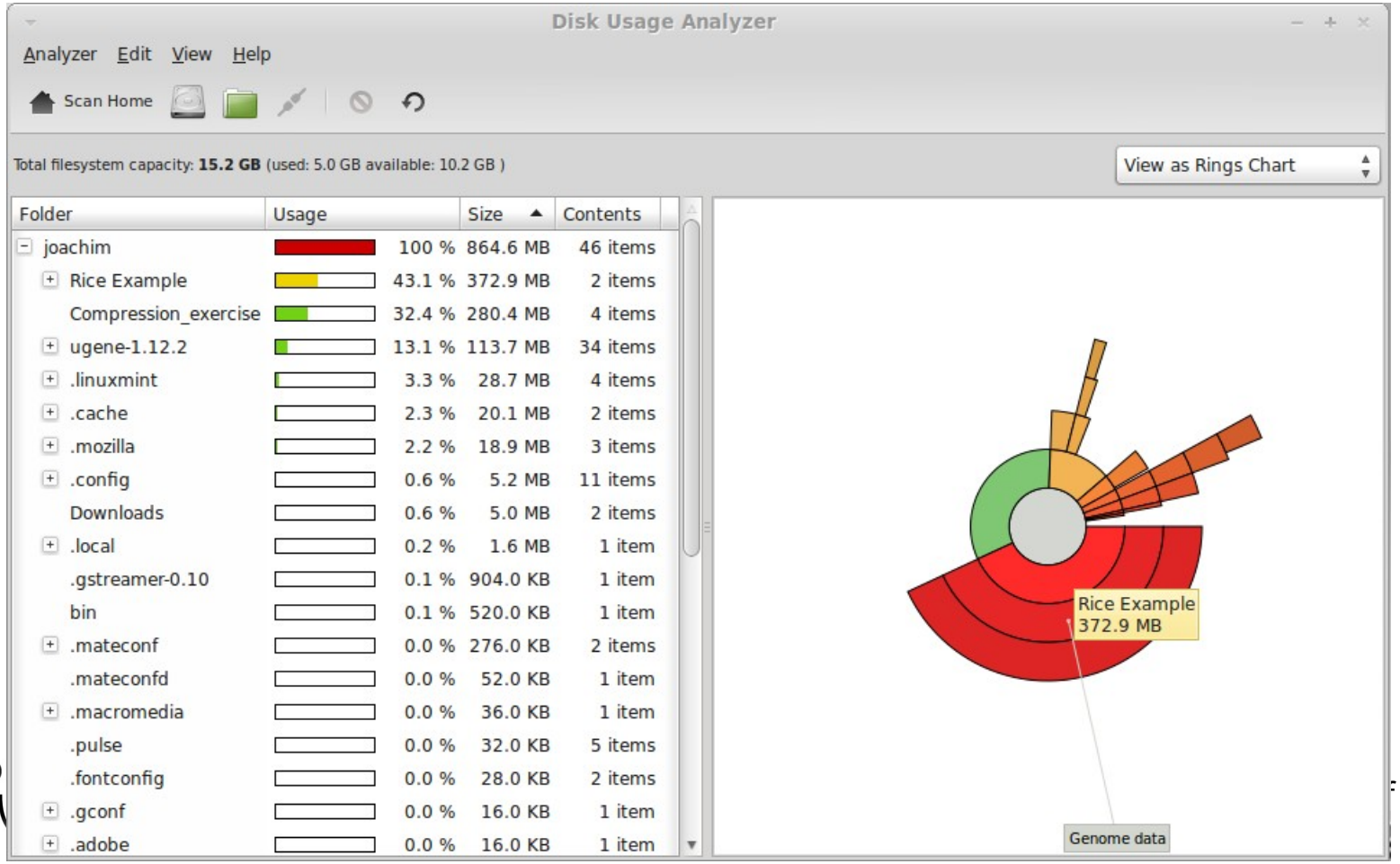
The program disk-utility put a lot of commands at work behind the scenes.

Some of the command line tools used:
- mount
- umount
- fdisk
- mkfs

You can read the man pages and search for guides on the internet if you want to get to know these (out of scope for this course).

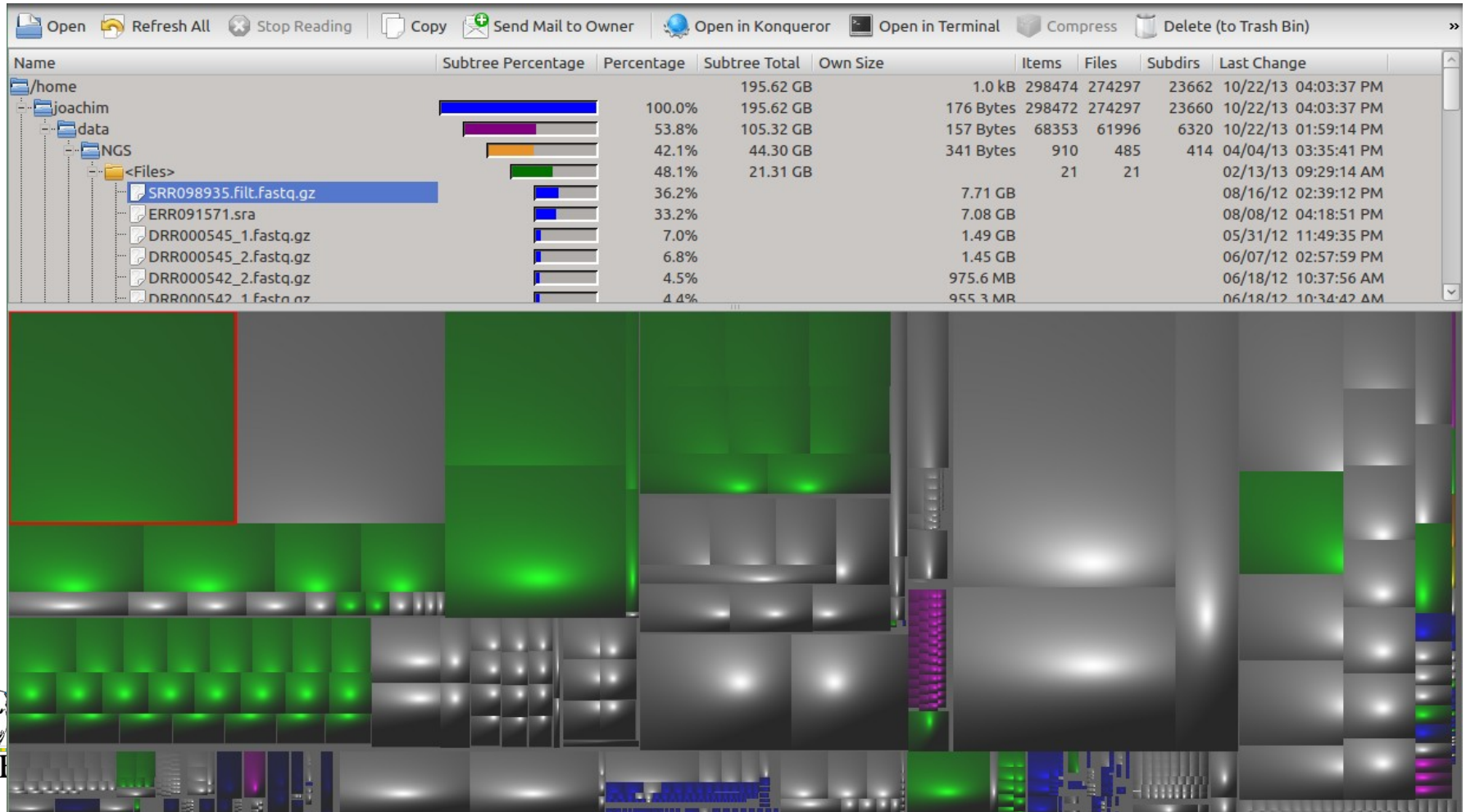
Checking storage space

By default 'disk usage analyzer'.



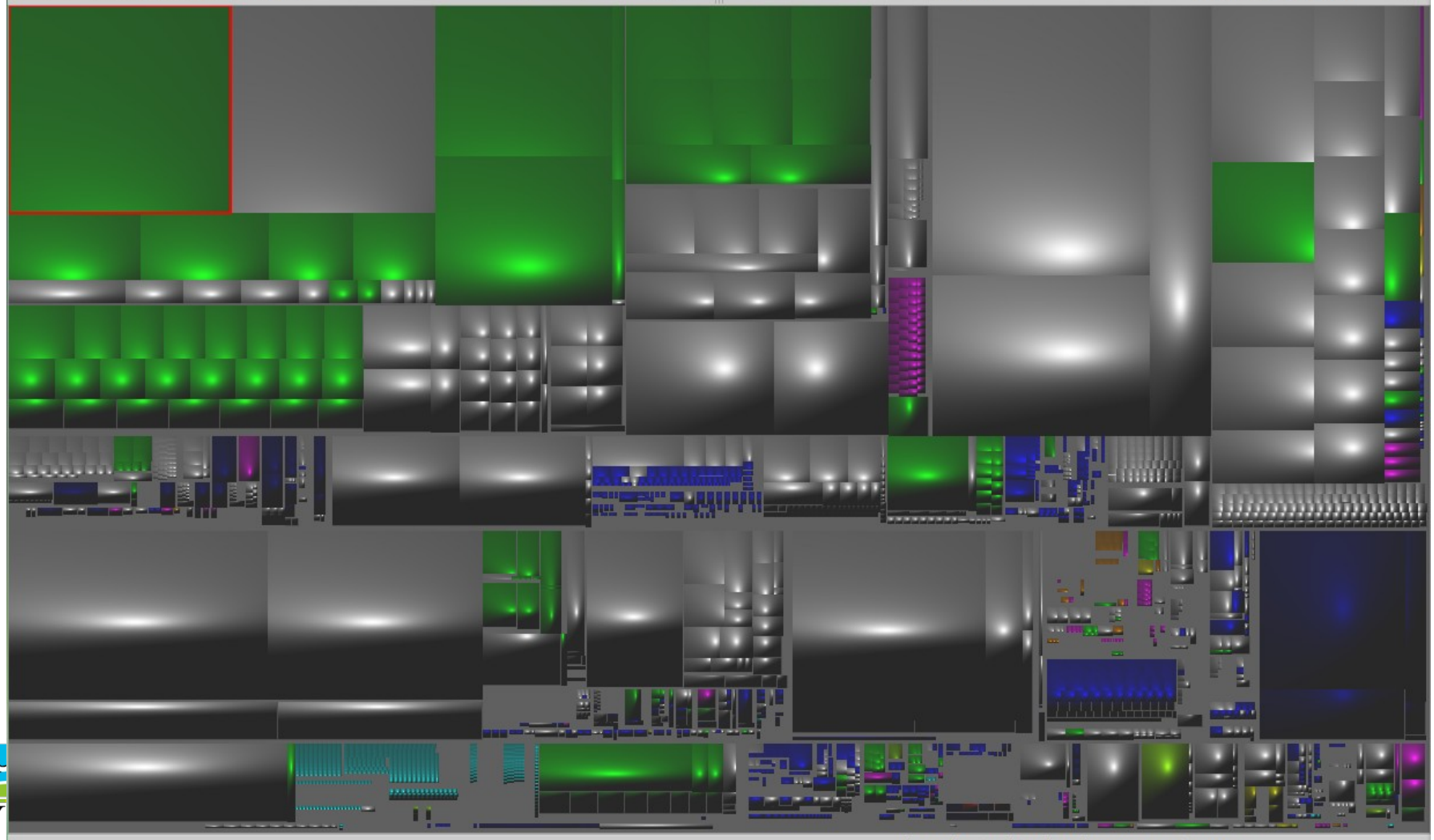
Checking storage space

Bonus: K4DirStat. Not installed by default.



Open | Refresh All | Stop Reading | Copy | Send Mail to Owner | Open in Konqueror | Open in Terminal | Compress | Delete (to Trash Bin)

Name	Subtree Percentage	Percentage	Subtree Total	Own Size	Items	Files	Subdirs	Last Change
/home			195.62 GB	1.0 kB	298474	274297	23662	10/22/13 04:03:37 PM
joachim		100.0%	195.62 GB	176 Bytes	298472	274297	23660	10/22/13 04:03:37 PM
data		53.8%	105.32 GB	157 Bytes	68353	61996	6320	10/22/13 01:59:14 PM
NGS		42.1%	44.30 GB	341 Bytes	910	485	414	04/04/13 03:35:41 PM
<Files>		48.1%	21.31 GB		21	21		02/13/13 09:29:14 AM
SRR098935.filt.fastq.gz		36.2%		7.71 GB				08/16/12 02:39:12 PM
ERR091571.sra		33.2%		7.08 GB				08/08/12 04:18:51 PM
DRR000545_1.fastq.gz		7.0%		1.49 GB				05/31/12 11:49:35 PM
DRR000545_2.fastq.gz		6.8%		1.45 GB				06/07/12 02:57:59 PM
DRR000542_2.fastq.gz		4.5%		975.6 MB				06/18/12 10:37:56 AM
DRR000542_1.fastq.gz		4.4%		955.3 MB				06/18/12 10:34:42 AM

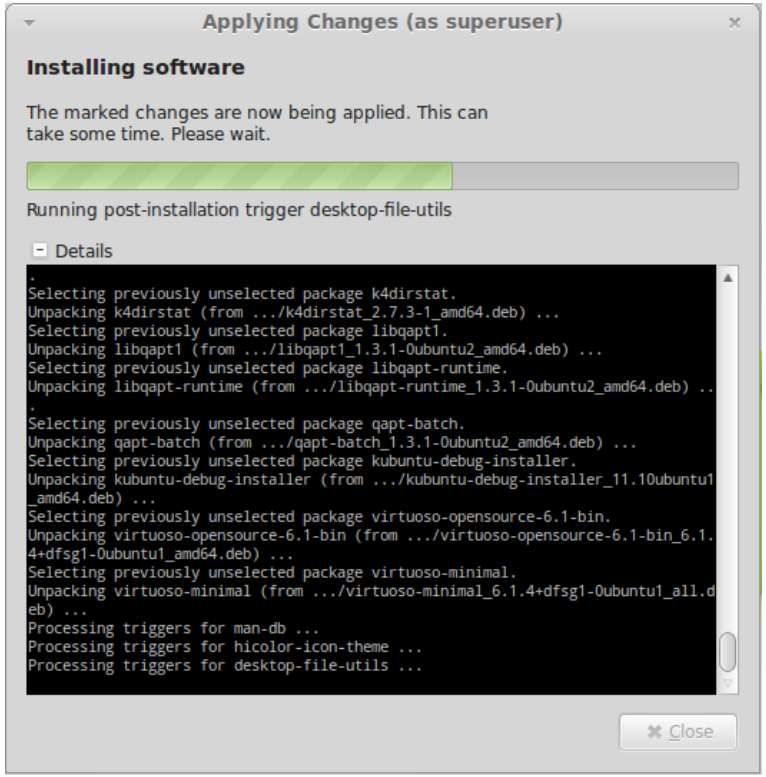


/home/joachim/data/NGS/SRR098935.filt.fastq.gz

K4Dirstat is a KDE package

Rehearsal: what is KDE?

Bonus: what happens when you install this package on our system?



VIB JAKONIX IRU

Space left on disks with df

To check the storage that is used on the different disks: **df -h**

```
~/ $ df -h
Filesystem      Size  Used Avail Use% Mounted on
/dev/sda1        12G  5.3G  5.7G  49% /
udev            490M  4.0K  490M   1% /dev
tmpfs           200M  920K  199M   1% /run
none            5.0M     0  5.0M   0% /run/lock
none            498M   76K  498M   1% /run/shm
/dev/sdb1       3.8G   20M  3.7G   1% /media/test

~/ $ df -h .
```

The size of directories

To check the size of files or directories: **du**

```
~/ $ du -sh *
520K   bin
281M   Compression_exercise
4.0K   Desktop
4.0K   Documents
5.0M   Downloads
4.0K   Music
4.0K   Pictures
4.0K   Public
373M   Rice Example
4.0K   Templates
4.0K   test
17Mtest.img
114M   ugene-1.12.2
4.0K   Videos
```

* means 'anything'.

This is called 'globbing':
* is a wild card symbol.

Wildcards on the command line

Wildcards are used to describe the names of **files/dirs**.

*

On that position, any length of string is allowed
e.g. **s*** matches: **s**an, **s**dd, **s**anitisation, **s**am.alignment,...

?

On that position, any character is allowed.
e.g. **saniti?ation** matches: saniti**s**ation, saniti**r**ation, ...

[ ]

On that position, the character may be one of the characters between [ ],
e.g. **saniti[sz]ation** matches: saniti**s**ation and saniti**z**ation

Wildcards on the command line

Many tools that require an **argument** to point to files or directories accept these wildcards.

```
~/ $ du -sh Do*
```

Wildcards on the command line

Many tools that require an **argument** to point to files or directories accept these wildcards.

```
~/ $ du -sh Do*
4.0K  Documents
20GDownloads
```

Wildcards on the command line

Many tools that require an **argument** to point to files or directories accept these wildcards.

```
~/ $ ls *.fastq
```

Wildcards on the command line

Many tools that require an **argument** to point to files or directories accept these wildcards.

```
~/ $ ls *.fastq
ERR148552_1.fastq                          ERR148552_2.fastq
testout.fastq
ERR148552_1_prinseq_good_zzwI.fastq  test.fastq
```

Keywords

Compression
Archive
Symbolic link
mounting
File system format
partition
Recursively
df
du
unlink

Write in your own words what the terms mean

Break