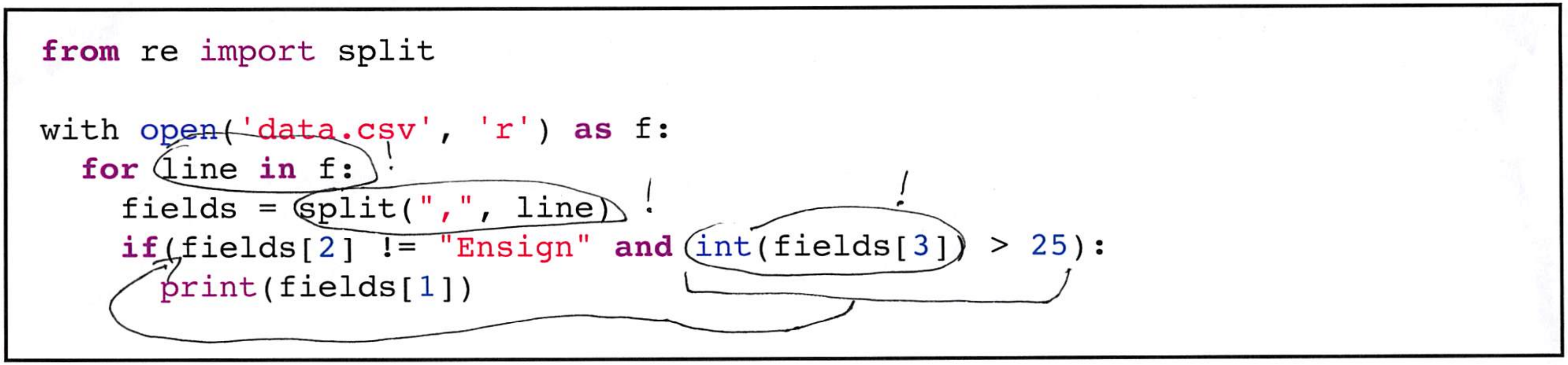
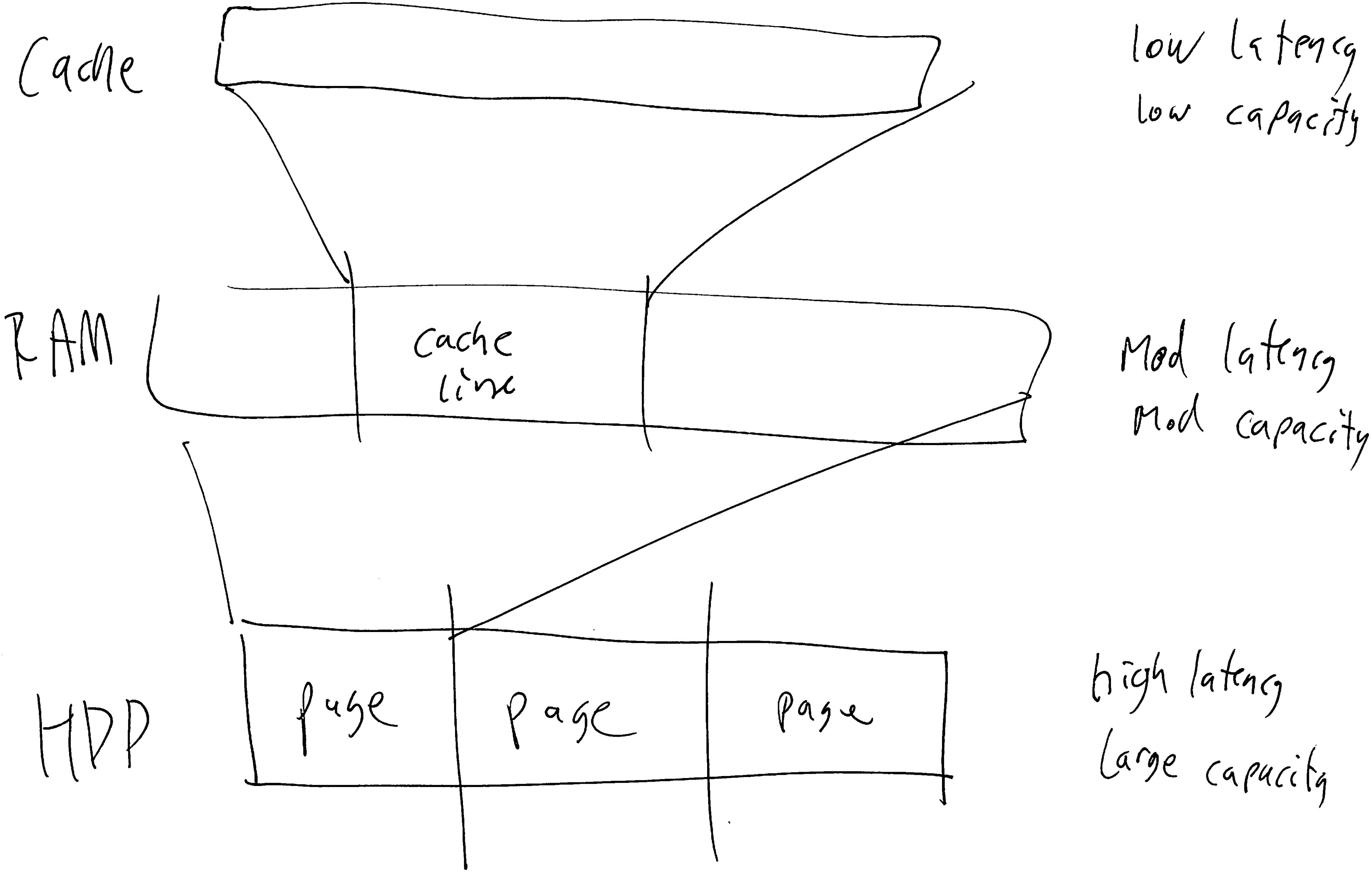
CSE 562

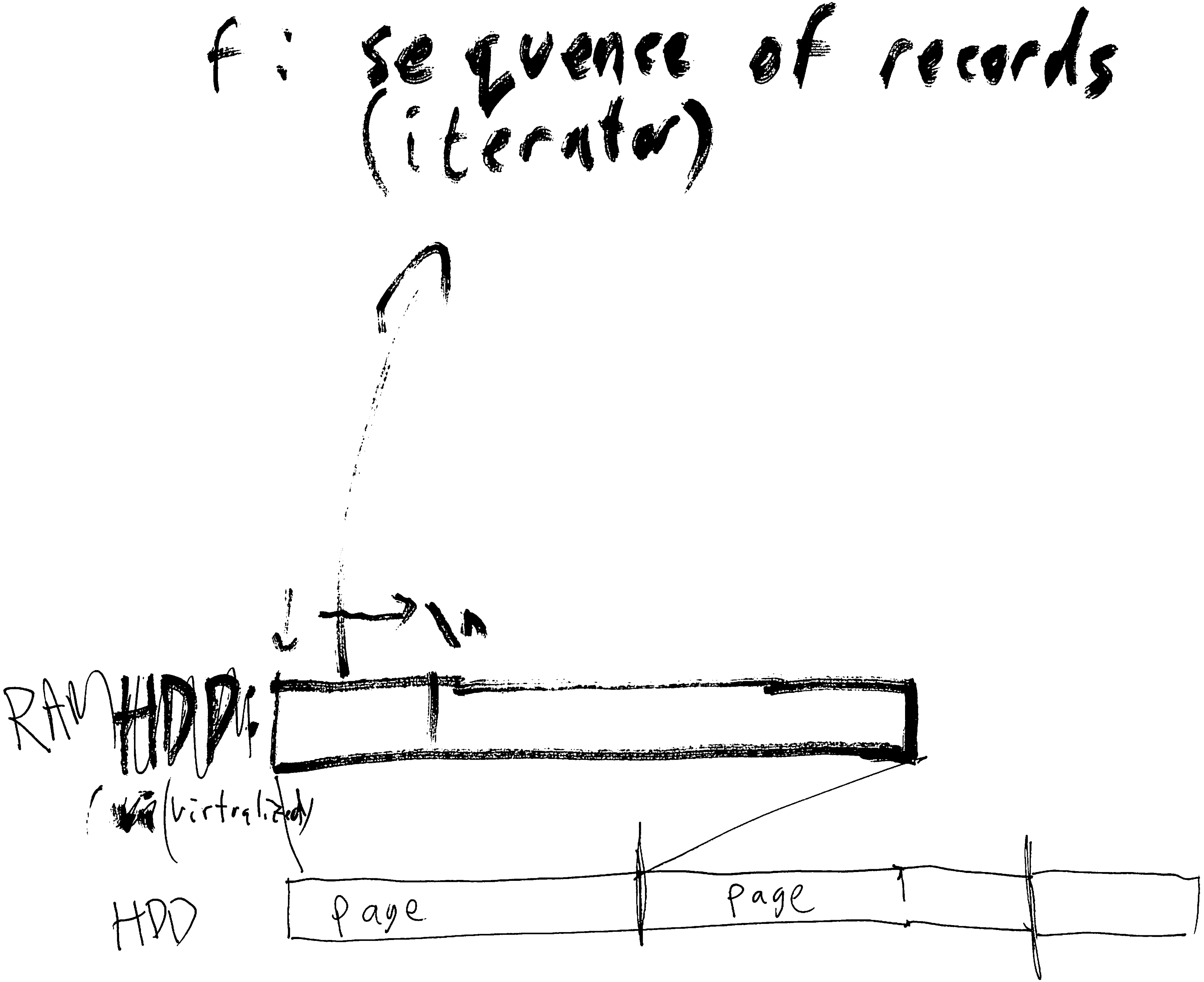
Physical data organization

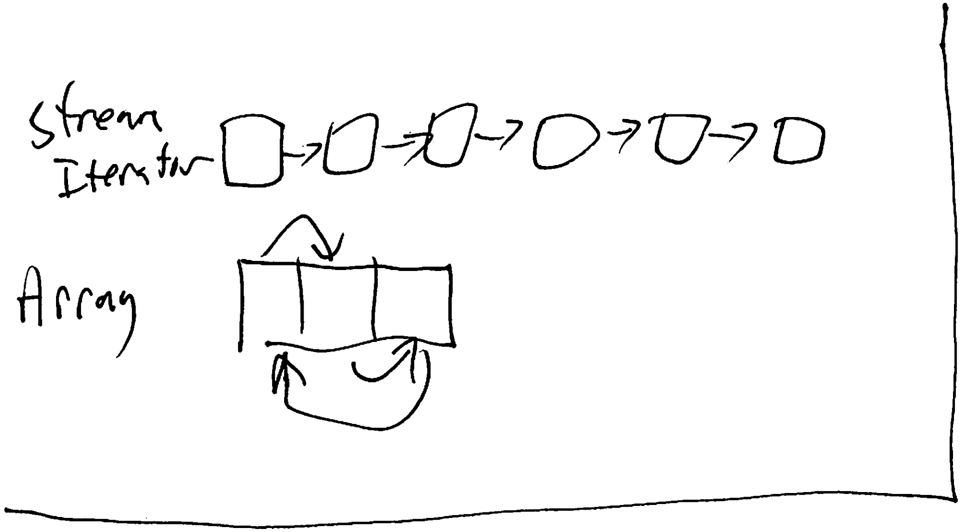


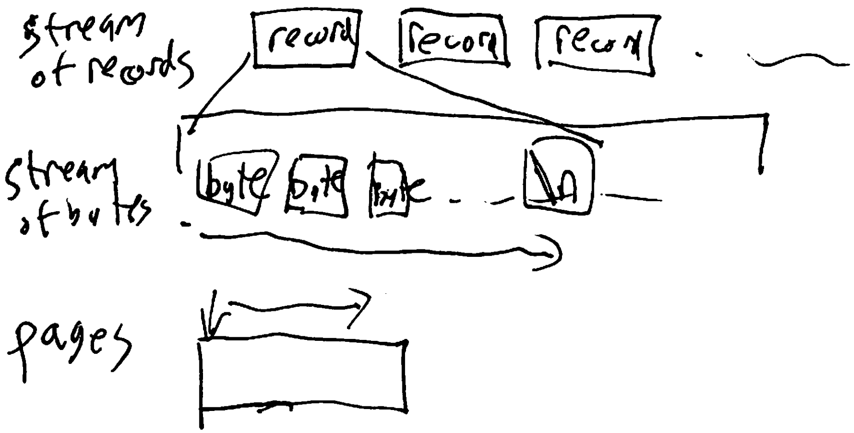
```
from re import split

with open('data.csv', 'r') as f:
  for line in f:
    fields = split(",", line)
    if(fields[2] != "Ensign" and int(fields[3]) > 25):
      print(fields[1])
```

```
1,Redshirt,Ensign,19,
2,Spock,Lt.,103,
3,Kirk,Capt.,22,
4,Redshirt,Ensign,21,
5,Redshirt,Ensign,18,
6,McCoy,Lt. Cmdr,38,
```


Cache
low latency
low capacity
RAM
cache
line
Mod latency
Mod capacity
HDP
page
page
page
high latency
large capacity


f : sequence of records
(iterator)
\n
RAM ~~HDD~~
(virtualized)
HDD
page
page

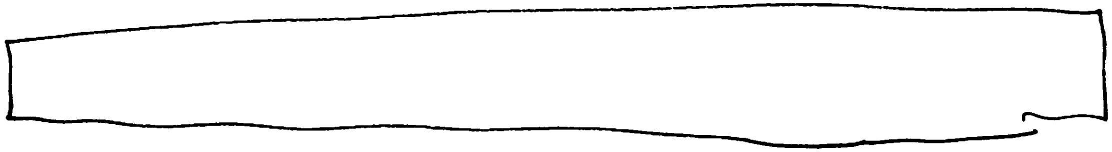

Stream Iterator
Array

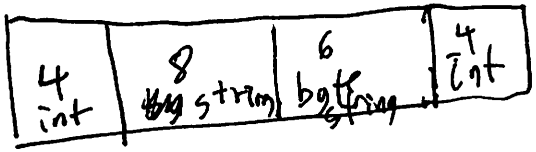

stream of records
record
record
record
stream of bytes
byte
byte
byte
\n
pages

Idea 1 store fields in a native format

pro: No casting

con: Need to save information about storage format

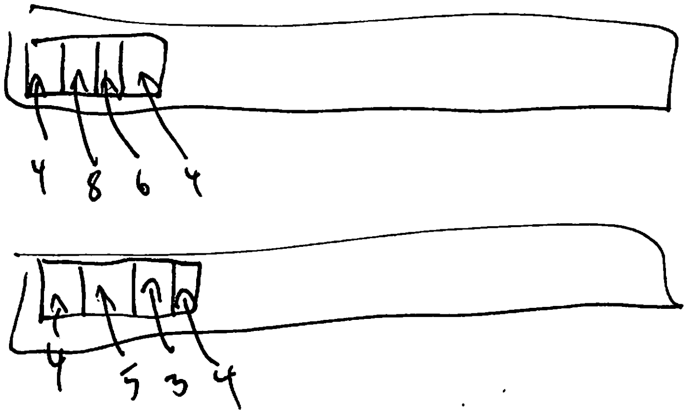




Fixed size encoding

pro: No delimiters (saves 4 bytes)

con: Need to know max length
Wasted space on variable length fields
~~Not~~ Not good for adding data



representing records

"CSV" → delimiters

"Fixed Width"

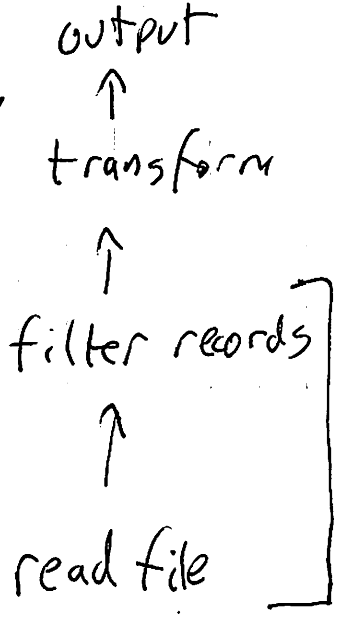
"Per record directory"

Factors

Size of the file

Types of the fields

How much data is changing


output
transform
filter records
read file