processando textos enormes com ferramentas "unix"

oi!

eu sou o luiz

me encontre em

@luiz_amf

github.com/lamenezes

o problema

abrir/ler e processar ~~textões~~ arquivos de texto com mais de ~~8 mil~~ **10 milhões** de linhas (>4 GB)

UNIX

Where there is a shell, there is a way.

*Write programs that **do one thing** and **do it well***

*Write programs to **work together**.*

*Write programs to **handle text streams**, because that is a **universal interface**.*

"Unix Philosophy" por Peter Salus

fazer uma coisa e fazer bem

a caixa de ferramentas

cat *con**cat**enate files and print*

- visualizar arquivos texto
- exemplo

```
$ cat foo.txt
```

less

- visualizar arquivos texto
- permite navegação
- lê arquivo enquanto executa
- exemplo

```
$ less foo.txt
```

cp *copy*

- sempre tenha um backup de seus dados
- exemplo

```
$ cp foo.txt backup.txt
```

head & tail

- imprime *X* linhas do arquivo
- dividir e conquistar
- exemplos

```
$ head foo.txt -n 20
$ tail foo.txt -n 50
```

WC *word count*

- qual o tamanho da bronca?
- calcula do arquivo
  - linhas
  - caracteres/bytes
  - palavras
- exemplo

```
$ wc foo.txt
```

cut

- remove partes de cada linha de um arquivo
- exemplo

```
$ cut -f2,3-5 foo.txt
```

sed *stream **ed**itor*

- editor "completo"
  - **substituição**/remoção de caracteres
  - duplica linhas
  - remoção de linhas
  - busca
- exemplo

```
$ sed 's/foo/bar/' foo.txt
```

grep *global search a regular expression and print*

- busca
- exemplo

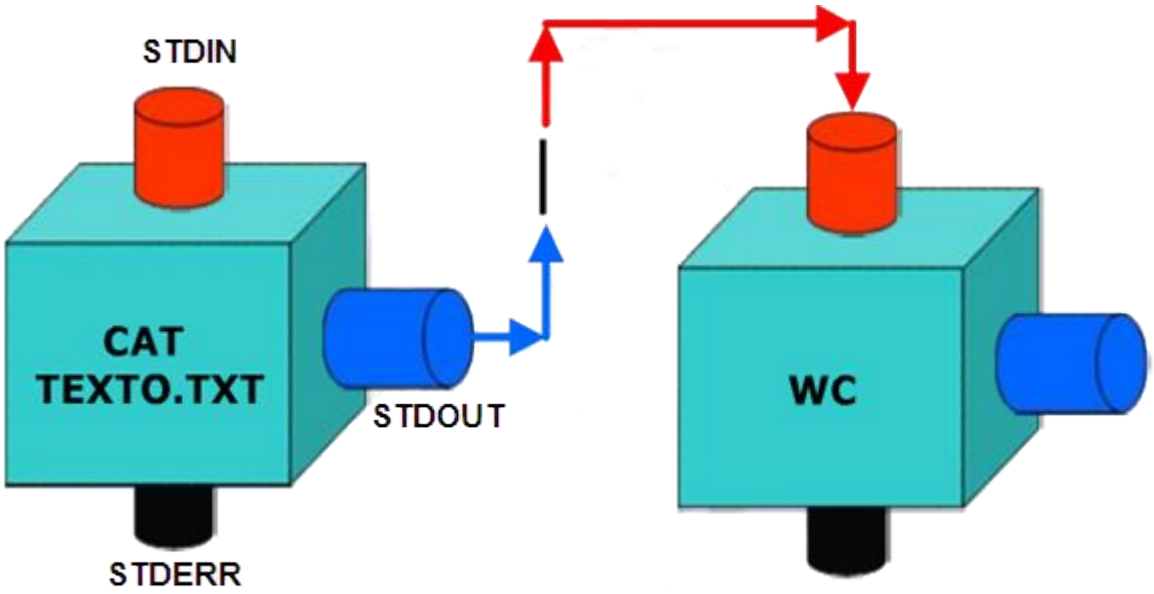
```
$ grep agulha palheiro.txt
```

trabalhar bem em conjunto

a interface universal de stream de textos

pipes

- encadeamento de comandos



tr *translate or delete characters*

```
cat <arquivo> | tr <de> <para>
```

- traduz caracteres
- deleta
- "aperta"

pipes

- busca com filtros múltiplos

```
cat random.csv | grep AC | grep "Sr\."
cat random.csv | grep João | grep AL
```

- busca + remoção de palavras

```
cat random.csv | grep AC | sed "s/Dr. //"
cat random.csv | grep GO | sed "s/Sr. //"
```

pipes

- visualizar consumo de memória dos programas

```
ps aux | sed "s/ \+/\t/g" | cut -f 4,11- | less
```

obrigado!

@luiz_amf
github.com/lamenezes

Credits

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival