FIELD NOTE / ARCH LINUX

Konversi PDF Menjadi PNG dengan GhostScript

Petruknisme · 10 Jan, 2016 · 3 min to read

Beberapa hari yang lalu saya berbincang di grup telegram tentang masalah OCR dan cara untuk ekstrak data dari PDF agar menjadi Image yang nantinya akan dipergunakan saat proses konversi dari image menjadi file teks agar lebih mudah dalam melakukan image recognition. Tidak semata-mata iseng hanya melakukan konversi pdf menjadi image, tapi memang sedang melakukan pengumpulan data yang nantinya akan dijadikan sebagai database.

Latar belakang itulah yang menjadikan saya ingin mencoba juga melakukan konversi pdf file menjadi image, sehingga satu halaman dalam file akan menjadi 1 file image yang sudah dikonversi. Awalnya, saya melakukan konversi menggunakan 'convert' yang dimiliki oleh ImageMagick. Dengan laptop berprosesor Intel i5, 4Core CPU dan dengan RAM 8GB saya kira cukuplah untuk melakukan pekerjaan berat seperti itu. Namun, apa yang saya sangka ternyata salah. Untuk proses melakukan konversi 3 halaman pdf menjadi 1200 dpi image file saja ImageMagick menghabiskan 5GB Ram dan 4 Core CPU. Rasanya seperti dicekik oleh ImageMagick langsung yang mengakibatkan laptop full hang dengan proses pemulihan cpu menjadi normal dibutuhkan 5menit agar menjadi stabil.

Saya putuskan bahwa ImageMagick tidaklah cocok dengan mesin yang saya miliki dan kebutuhan yang sedang dikerjakan. Salah satu anggota grup telegram tersebut memberikan saran bahwa GhostScript yang akan saya pakai dalam artikel ini lebihlah ringan dan tidak sekejam ImageMagick dalam melakukan proses yang sudah saya sebutkan sebelumnya. File yang saya lakukan konversi mempunyai 1490 halaman dan hanya menghabiskan waktu 9015 detik dalam melakukan konversi dengan 1GB hasil yang didapatkan.

Pada artikel ini saya tidak bisa menggunakan file pdf yang diceritakan karna file tersebut bersifat pribadi dan belum bisa saya publikasikan. Sebagai gantinya saya menggunakan file pdf berjudul CodeIgniter Documentation 3.0-dev yang bisa didownload pada link berikut : https://media.readthedocs.org/pdf/codeigniter-30-dev/readthedocs/codeigniter-30-dev.pdf (https://media.readthedocs.org/pdf/codeigniter-30-dev/readthedocs/codeigniter-30-dev.pdf) . File tersebut mempunyai 685 halaman yang dimana lebih sedikit dibandingkan dengan yang saya ceritakan sebelumnya.

Kita lakukan instalasi GhostScript terlebih dahulu sebelum memulai melakukan konversi. Disini saya menggunakan OS Arch Linux sebagai distro yang saya pakai. Karna bisa jadi pembaca artikel ini menggunakan distro yang lain maka saya akan lampirkan referensi atau cara instalasi untuk distro selain Arch Linux. Untuk Arch Linux sendiri GhostScript sudahlah ada pada repo official jadi kita hanya tinggal melakukan instalasi lewat **pacman** seperti berikut :

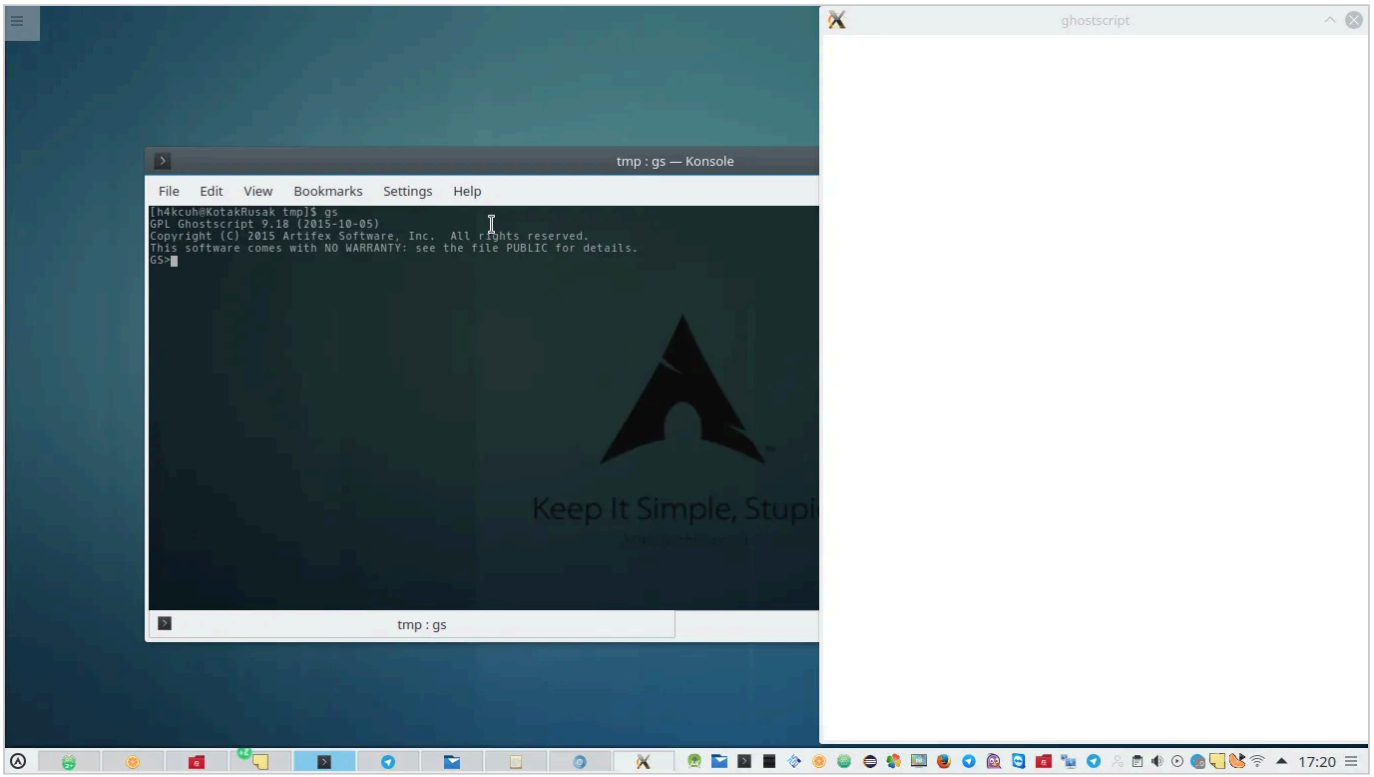
```
1 Pacman -Syu
```

Licensed under CC BY-SA 4.0

```
2
3   Pacman -S ghostscript
```

Untuk distro lain bisa melihat pada referensi yang diberikan oleh ghostscript yang bisa dilihat pada : **http://ghostscript.com/doc/current/Install.htm** (http://ghostscript.com/doc/current/Install.htm) . Dan silahkan disesuaikan dengan distro masing-masing yang mungkin saja sudah ada paket dalam repo ofisial distronya.

Saya asumsikan bahwa kita sudah melakukan instalasi untuk GhostScript pada sistem yang kita miliki. Agar tahu bahwa ghostscript sudah berhasil terinstal adalah dengan cara mengetikkan 'gs' pada terminal/konsole yang nantinya akan muncul tampilan seperti berikut:



Setelah kita tahu bahwa GhostScript sudah terinstal dan menunjukkan versi 'GPL Ghostscript 9.18 (2015-10-05)' maka selanjutnya ketik 'quit' untuk keluar dari konsol 'gs' karena bukan itu yang akan kita pakai. Selanjutnya kita lakukan perintah berikut di terminal/konsole :

```
1   mkdir imageresult
2   nano Ghost_script_pdf_to_image.sh
```

Tambahkan bash script seperti berikut pada file *Ghost_script_pdf_to_image.sh :*

Ghost_script_pdf_to_image.sh

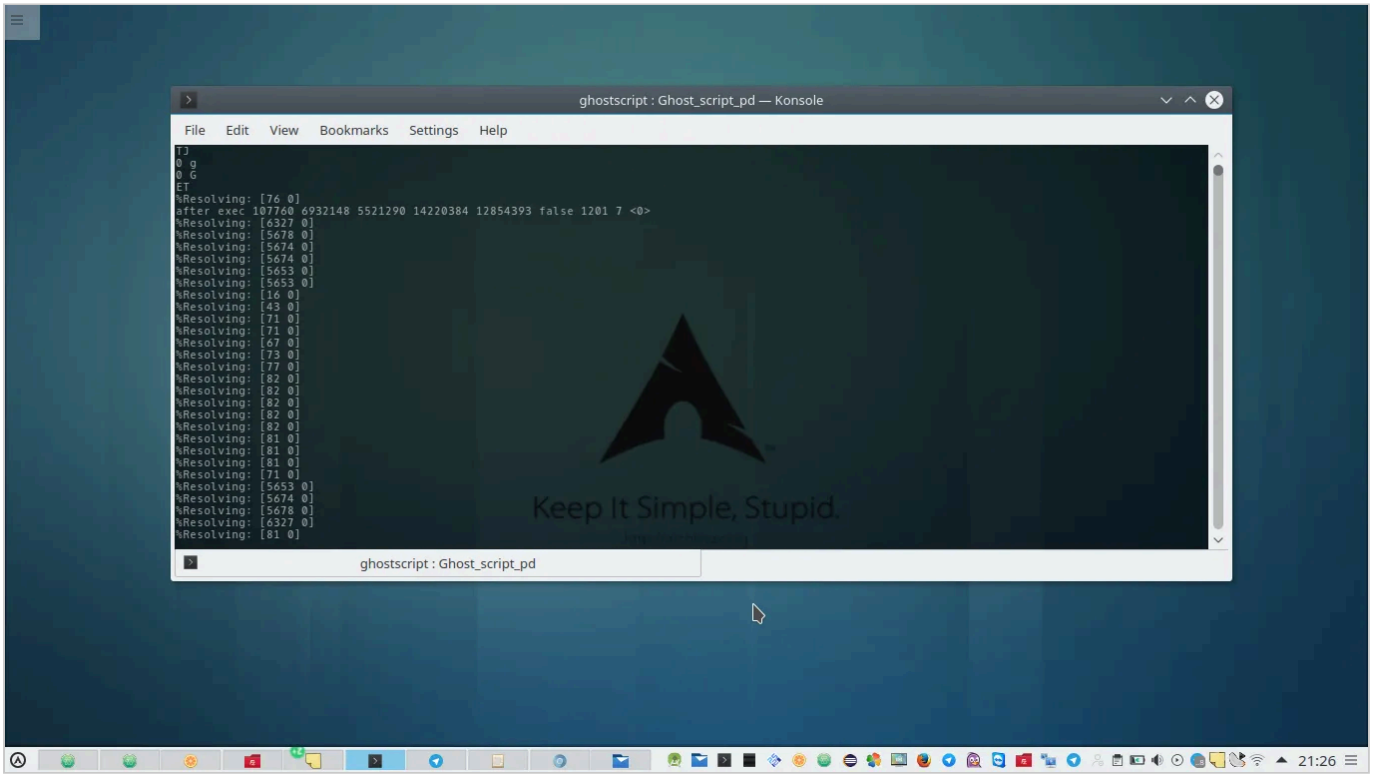
```
#!/bin/bash
START=$(date +%s)
pageNum=`gs -q -dNODISPLAY -c "(./codeigniter-30-dev.pdf) (r) file runpdfb
gs -dDEBUG -dNumRenderingThreads=4 -dNOPAUSE -sDEVICE=pngalpha -dFirstPage

END=$(date +%s)
DIFF=$(( $END - $START ))
echo "It took $DIFF seconds"
```

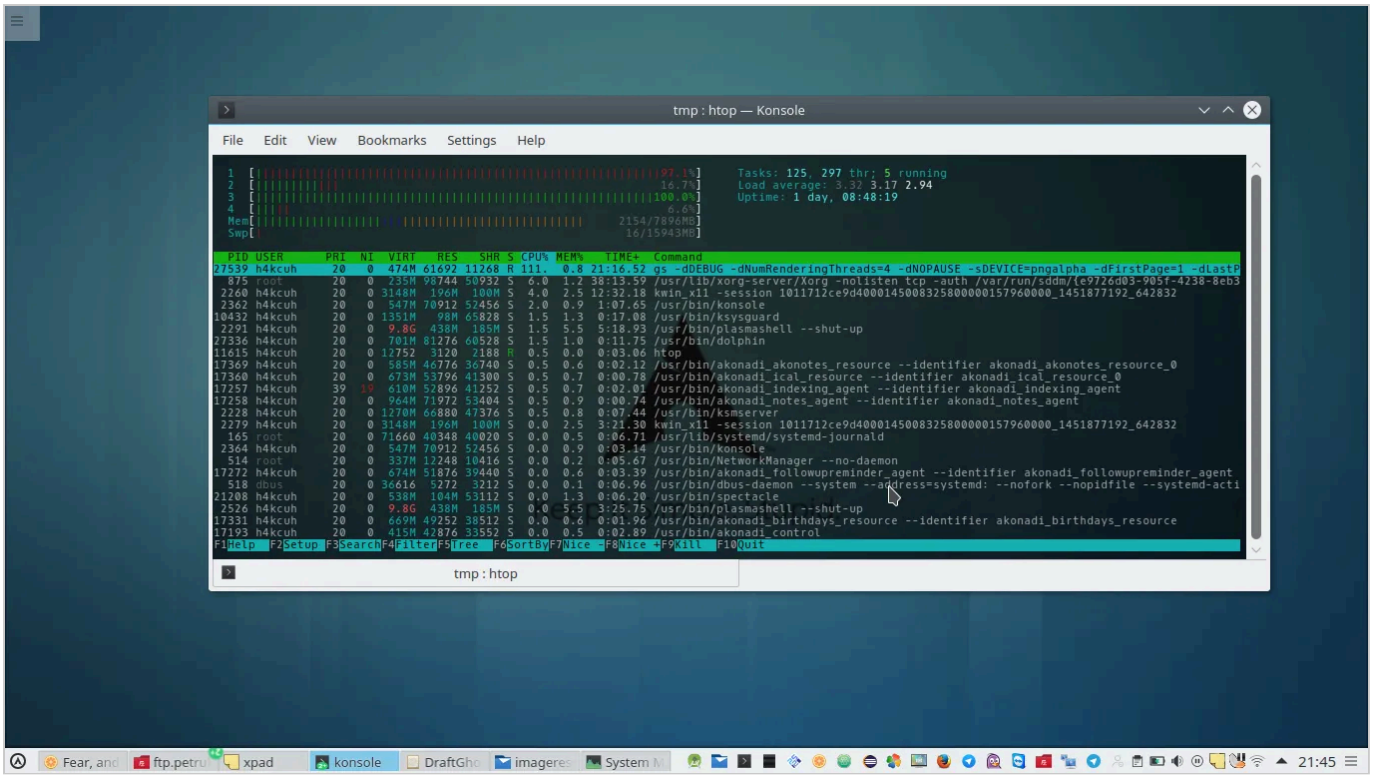
Setelah itu save file dan kemudian berikan akses untuk melakukan eksekusi seperti berikut :

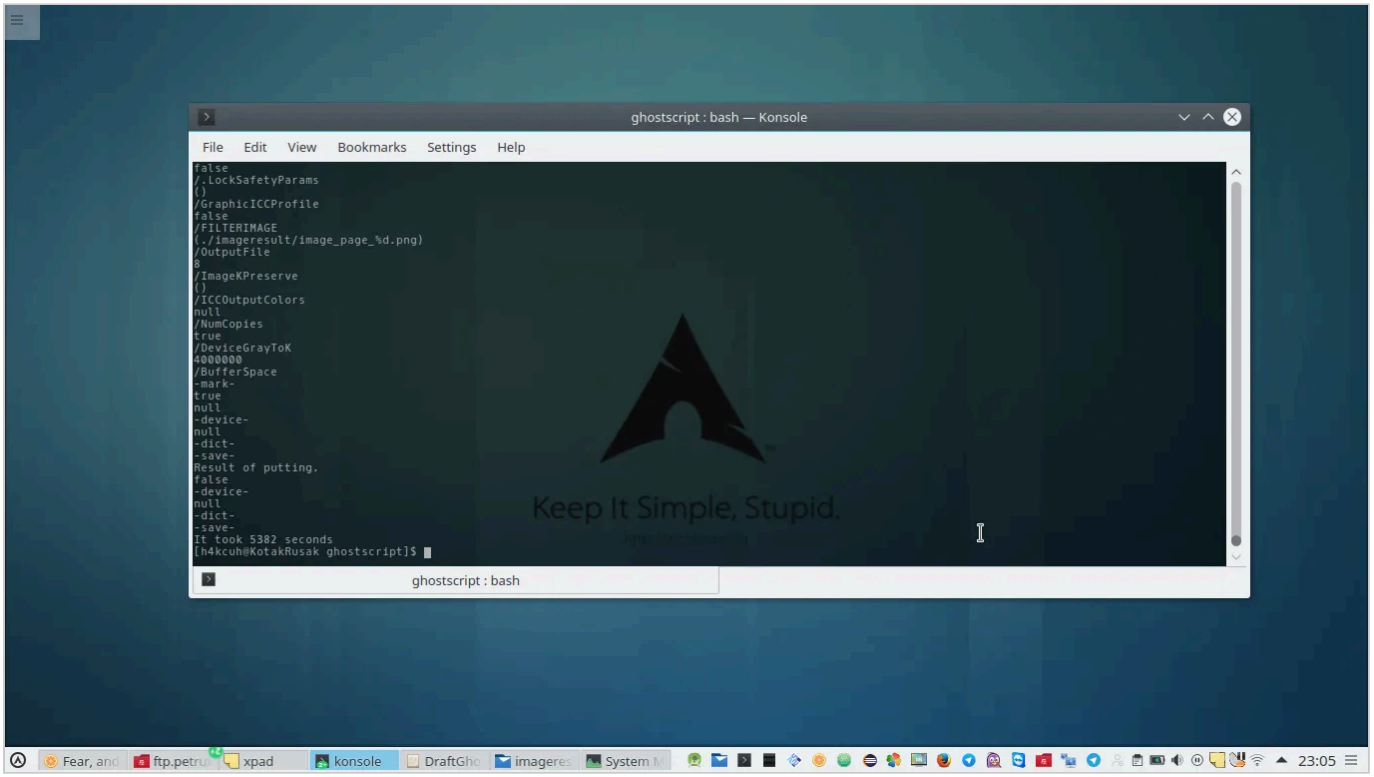
```
chmod +x Ghost_script_pdf_to_image.sh
```

Lalu jalankan bash script yang sudah kita buat dan akan tampil debug output seperti berikut :

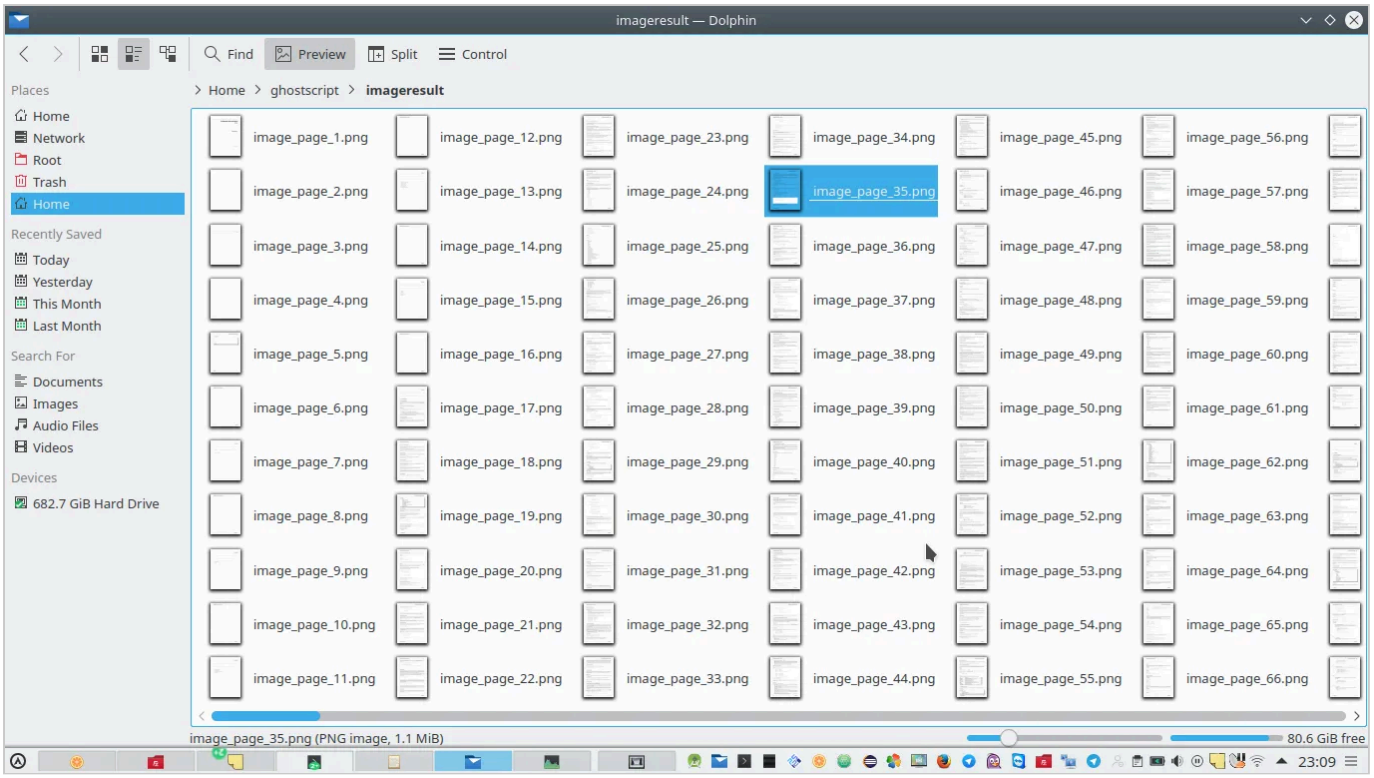


Lamanya waktu saat konversi itu bergantung pada mesin yang anda miliki dan pada jumlah yang anda proses. Berikut saya tampilkan beberapa screenshot saat melakukan proses konversi sampai dengan selesai.





Licensed under CC BY-SA 4.0



Kita sudah melakukan konversi PDF menjadi PNG dengan density 1200dpi dan menghasilkan 685 file, 5382 detik, 791.7 MB. Dan kesimpulan yang bisa diambil adalah GhostScript lebihlah ringan dan powerfull dibandingkan dengan ImageMagick yang memerlukan resource yang sangat besar tapi itu semua bergantung kepada kebutuhan yang dimiliki oleh masing-masing. Akhir kata, saya berharap artikel ini bisa bermanfaat.

Arch Linux # Tutorial

Licensed under CC BY-SA 4.0