VOLKSWAGEN

AKTIENGESELLSCHAFT

Developing a Cloud Computing Based Approach for Forensic Analysis using OCR

Matthias Trojahn, Volkswagen AG, Germany
Lei Pan, Deakin University, School of IT, Australia
Fabian Schmidt, ISC Gebhardt, Germany

Outline

1. Motivation
2. Background
3. Developing a Cloud-based Framework
4. Implementing Our System
5. Conclusion

Motivation

- Digital forensic tools to extract information
- Criminals: Information hiding in files
- Embedding screenshot of text in PDF files

↓

This text could be extracted!

Processing different file types

	Before OCR	With OCR
→	Plain text of the file	Text and images of the file
→	Plain text of the file	Text and images of the file
→	Plain text of the file	Plain text of the file
→	nothing	Information of the picture
→	nothing	nothing

Advantage Optical Character Recognition (OCR)

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

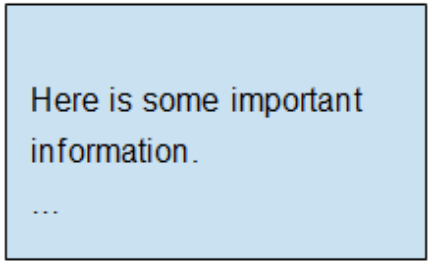


Figure 1: Example figure

Lorem ipsum dolorsit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet,

→

dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Here is some important information.

…

Figure 1: Example figure

With OCR

Lorem ipsum dolorsit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos

- Embedded in standalone tools like FTK (with Version 3.1)
- Performance is the bottleneck
- Problems with system scalability

Our aims

1) Cloud-based approach to perform OCR jobs for forensic purposes;
2) Clearly defined communication protocol between the job controller and each virtual machine;
3) Analytical solution for setting up multiple virtual machines

Challenges

1) Most cloud-based systems do not naturally support OCR
2) Standard operating systems available on cloud are unnecessarily large
3) Current cloud-based solutions lack a centralized management tool to organize the jobs

Deriving the number of virtual machines

Probability of each incoming job is delayed

$$P_d = \frac{P_0 (m\rho)^m}{m!(1-\rho)}$$

with

$$P_0 = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!\,(1-\rho)} \right]^{-1}$$

Time to complete a job

$$T = \frac{P_d}{m\mu - \lambda} + \frac{1}{\mu}$$

p: utilization rate $\rho = \frac{\lambda}{m\mu} < 1$

m: number of VMs

μ: average engaging rate of each VM

λ: average incoming rate of each job

Example for the calculation

Convert 10 pages in 100 seconds

$\mu = 10/100 = 0.1$

Process 12 pages per minute

$\lambda = 12/60 = 0.2$

Probability of delay (for three VMs)

$P_d = 0.2353$

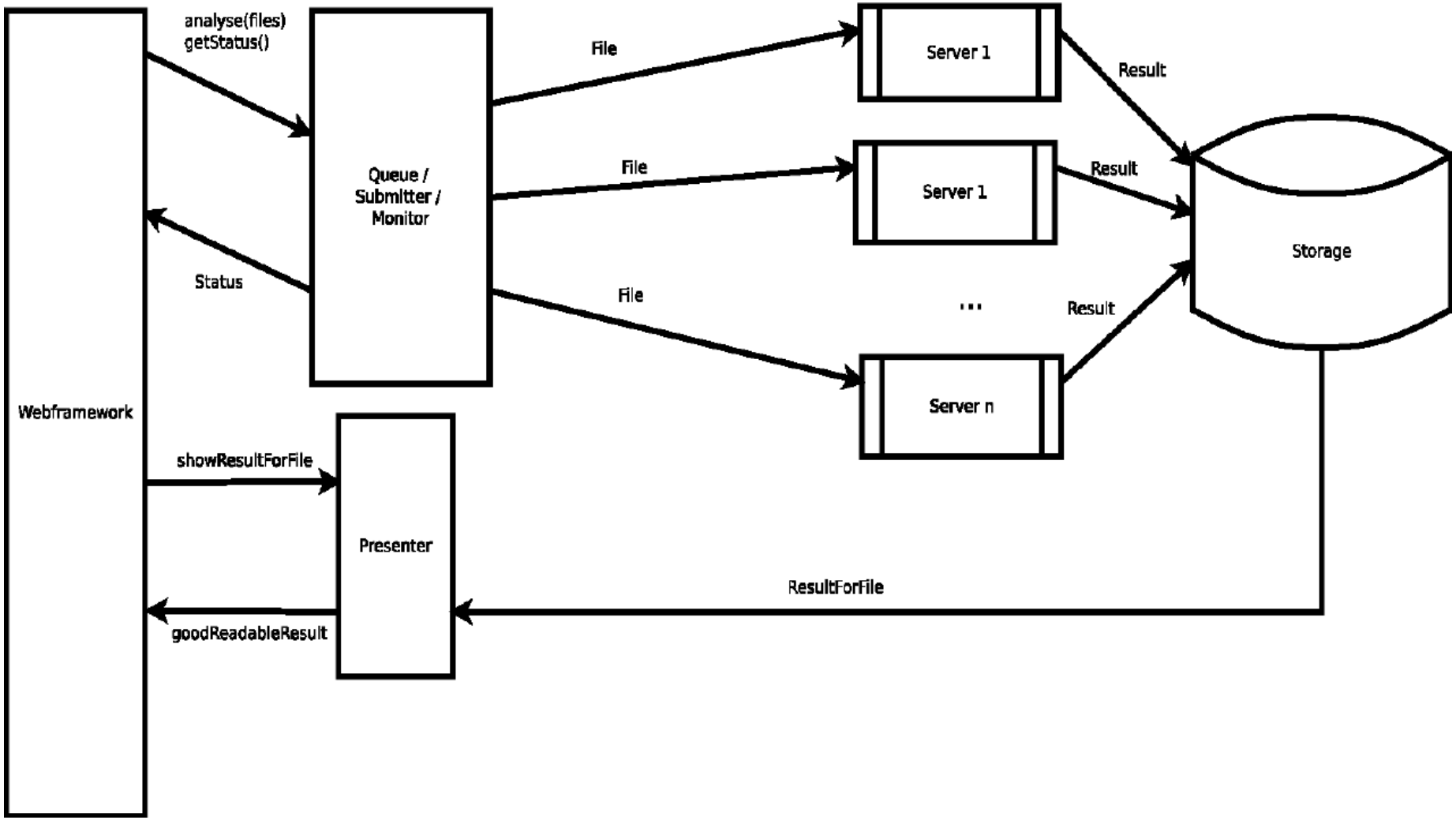
<u>Result:</u>

$T = 12.353$ sec → Processing 17,280 PDF pages per day

Why processing the whole page?

- No context information get lost

Implementing Our System - Architecture



Implementing Our System

Forensic tool

Please upload a pdf file or a zip-file

select

Start to compute

Currently working processes

id	hash	status	download
1	zna7M...	finished	download
2	KwcN1...	finished	download
3	BFUQe...	running	-
4	RXjRD...	running	-
5	XpTz9...	running	-
6	n94gl...	running	-
7	8fklZ...	not started	-

ImageMagick

“convert -sharpen 2 -type Grayscale" + density + "-units PixelsPerInch - depth 8" + input + " " + tempfile + "%d.tif“

Tesseract (for all files)

“tesseract " + file + str(i) + “.txt >>"+output

Conclusion

- Proposed a cloud based system for forensic analysis (open-source)
- Scalable system (cloud based system using virtual machines)
- Framework improves scalability, performance, flexibility and upgradability

- Future studies on breaking CAPTCHA by using our OCR

Thank you for your attention!

Contact:

Matthias Trojahn

Matthias.Trojahn@volkswagen.de

+49 1520 1658545

Developing a waiting policy

- All pages have to be processed before showing

Probability of saturation the system with *s* jobs:

$$P_S = \frac{\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^S}{1-\left(\frac{\lambda}{\mu}\right)^{S+1}} = 0.0009837$$

S=5

Number of blocked jobs:

$$N = \sum_{k=0}^{S} kP_k = \frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}} - \frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}}(S+1)P_S = 0.6615$$

Time that a job spends in the system:

$$T = \frac{N}{\lambda(1-P_S)} = 3.3105 \text{ sec}$$

More jobs then machines

$$m\mu P_k = \lambda P_{k-1}$$

$$\begin{aligned} P_k &= \frac{\lambda}{m\mu} P_{k-1} \\ &= \left(\frac{\lambda}{\mu}\right)^{k-m} \left(\frac{1}{m}\right)^{k-m} P_m \\ &= \frac{(m\rho)^{k-m}}{m^{k-m}} \left(\frac{m\rho^m}{m!} P_0\right) \\ &= \frac{m^m \rho^k}{m!} P_0 \end{aligned}$$