2022 IEEE 15th International Conference on Cloud Computing (CLOUD)

Delivering Document Conversion as a Cloud Service with High Throughput and Responsiveness

1st Christoph Auer
IBM Research
Rüschlikon, Switzerland
cau@zurich.ibm.com

2nd Michele Dolfi
IBM Research
Rüschlikon, Switzerland
dol@zurich.ibm.com

3rd André Carvalho
SoftINSA Lda.
Tomar, Portugal
afecarvalho@gmail.com

4th Cesar Berrospi Ramis
IBM Research
Rüschlikon, Switzerland
ceb@zurich.ibm.com

5th Peter W.J. Staar
IBM Research
Rüschlikon, Switzerland
taa@zurich.ibm.com

2022 IEEE 15th International Conference on Cloud Computing (CLOUD) | 978-1-6654-8137-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/CLOUD55607.2022.00060

***Abstract*—Document understanding is a key business process in the data-driven economy since documents are central to knowledge discovery and business insights. Converting documents into a machine-processable format is a particular challenge here due to their huge variability in formats and complex structure. Accordingly, many algorithms and machine-learning methods emerged to solve particular tasks such as Optical Character Recognition (OCR), layout analysis, table-structure recovery, figure understanding, etc. We observe the adoption of such methods in document understanding solutions offered by all major cloud providers. Yet, publications outlining how such services are designed and optimized to scale in the cloud are scarce. In this paper, we focus on the case of document conversion to illustrate the particular challenges of scaling a complex data processing pipeline with a strong reliance on machine-learning methods on cloud infrastructure. Our key objective is to achieve high scalability and responsiveness for different workload profiles in a well-defined resource budget. We outline the requirements, design, and implementation choices of our document conversion service and reflect on the challenges we faced. Evidence for the scaling behavior and resource efficiency is provided for two alternative workload distribution strategies and deployment configurations. Our best-performing method achieves sustained throughput of over one million PDF pages per hour on 3072 CPU cores across 192 nodes.**

***Index Terms*—cloud applications, document understanding, distributed computing, artificial intelligence**

I. Introduction

Over the past decade, many organizations have accelerated their transformation into data-driven businesses, as studies have shown its positive impact in efficiency, decision making, or financial performance [1], [2]. Leading companies are increasingly deploying workloads on public and private cloud infrastructure, including business intelligence processing and machine learning models in data analytics platforms [3]. This is owed to several factors such as high availability, lower cost for compute, and storage [4], as well as the flexibility to scale up or down a cloud-based business process to fit the operational needs. Workloads and services can be containerized, deployed, and orchestrated through widely adopted and standardized platforms like Kubernetes [5], [6].

A key business process relevant to many companies is document understanding. Documents may constitute contracts, guidelines, manuals, presentations, papers, etc., which contain valuable knowledge for their operations. We observe that several specialized companies and all major cloud providers offer dedicated services (SaaS) for various aspects of document understanding such as Optical Character Recognition (OCR) (e.g., Amazon Textract[1]), forms, and invoice parsing (Docparser[2], Nanonets[3], Google Document AI[4], Microsoft SharePoint Syntex[5]), or conversion of unstructured formats such as PDF into structured content (IBM Watson Discovery[6]).

Conversion of PDF documents into a structured, machine-processable format is a particularly challenging business process due to the high variability and weak normalization of its input. To name a few dimensions of variability, PDF documents can be short or long, encode programmatic or scanned content, have simple or complex page layouts, may contain tables or figures, etc. Thus, the process of recovering their structure and extracting content in high detail entails several dynamic steps (see Fig. 1). On the computational side, this relies on multiple algorithms and machine-learning (ML) models specialized for particular tasks. Examples for such models include OCR [7], document layout analysis [8]–[10], table structure recovery [11], [12], figure understanding [13], reference and citation resolution [14], etc. Furthermore, the ML landscape is evolving rapidly, with new models frequently exposing significantly different characteristics in terms of computational expenses, memory usage, or accelerator re-

[1]https://aws.amazon.com/textract
[2]https://docparser.com
[3]https://nanonets.com/invoice-ocr
[4]https://cloud.google.com/document-ai
[5]https://docs.microsoft.com/en-us/microsoft-365/contentunderstanding
[6]https://www.ibm.com/cloud/watson-discovery

2159-6190/22/$31.00 ©2022 IEEE
DOI 10.1109/CLOUD55607.2022.00060