

**INTEGRASI TEKNOLOGI OCR DENGAN PROSES ETL
UNTUK OTOMATISASI DATA PIPELINE ARSIP PENCAIRAN
KEUANGAN BPS KABUPATEN SUKABUMI**

SKRIPSI

MUHAMMAD RAIHAN IZHARUL HAQ

20210040075



**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNIK, KOMPUTER, DAN DESAIN
UNIVERSITAS NUSA PUTRA
SUKABUMI
JUNI 2025**

**INTEGRASI TEKNOLOGI OCR DENGAN PROSES ETL
UNTUK OTOMATISASI DATA PIPELINE ARSIP PENCAIRAN
KEUANGAN BPS KABUPATEN SUKABUMI**

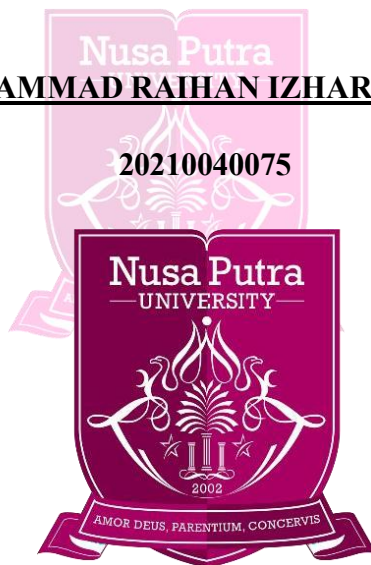
SKRIPSI

Diajukan Untuk Memenuhi Salah Satu Syarat Dalam Menempuh

Gelar Sarjana Teknik Informatika

MUHAMMAD RAIHAN IZHARUL HAQ

20210040075



**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNIK, KOMPUTER, DAN DESAIN
UNIVERSITAS NUSA PUTRA
SUKABUMI**

JUNI 2025

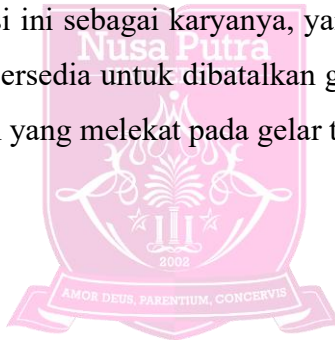
PERNYATAAN PENULIS

JUDUL : INTEGRASI TEKNOLOGI OCR DENGAN PROSES ETL
UNTUK OTOMATISASI DATA PIPELINE ARSIP
PENCAIRAN KEUANGAN BPS KABUPATEN SUKABUMI

NAMA : MUHAMMAD RAIHAN IZHARUL HAQ

NIM 20210040075

“Saya menyatakan dan bertanggungjawab dengan sebenarnya bahwa Skripsi ini adalah hasil karya sendiri kecuali cuplikan dan ringkasan yang masing-masing telah dijelaskan sumbernya. Jika pada waktu selanjutnya ada pihak lain yang mengklaim bahwa Skripsi ini sebagai karyanya, yang disertai dengan bukti-bukti yang cukup, maka saya bersedia untuk dibatalkan gelar Sarjana Komputer beserta segala hak dan kewajiban yang melekat pada gelar tersebut.”



Sukabumi, 30 Juni 2025

MUHAMMAD RAIHAN IZHARUL HAQ

Penulis

PENGESAHAN SKRIPSI

JUDUL : INTEGRASI TEKNOLOGI OCR DENGAN PROSES ETL
UNTUK OTOMATISASI DATA PIPELINE ARSIP
PENCAIRAN KEUANGAN BPS KABUPATEN SUKABUMI

NAMA : MUHAMMAD RAIHAN IZHARUL HAQ
NIM 20210040075

Skripsi ini telah diujikan dan dipertahankan di depan Dewan Penguji pada Sidang Skripsi tanggal 18 Juni 2025 Menurut pandangan kami, Skripsi ini memadai dari segi kualitas untuk tujuan penganugerahan gelar Sarjana Komputer (S.Kom).

Sukabumi, 30 Juni 2025

Pembimbing I

Pembimbing II

Gina Purnama Insany, S.Si.T., M.Kom

NIDN. 0417077908

Ir. Somantri, S.T., M.Kom

NIDN. 0419128801

Ketua Penguji

Ketua Program Studi Teknik Informatika

Indra Yustiana, S.T., M.Kom

NIDN. 0409017604

Ir. Somantri, S.T., M.Kom

NIDN. 0419128801

Plh.Dekan Fakultas Teknik Komputer dan Desain

Ir. Paikun, S.T., M.T., IPM., ASEAN Eng

NIDN. 0402037401

ABSTRACT

In the digital era, managing archival data poses challenges for many institutions, including the Central Statistics Agency (BPS) of Sukabumi Regency, especially when dealing with unstructured PDF documents. This study develops a data pipeline by effectively integrating Optical Character Recognition (OCR) technology and the Extract, Transform, Load (ETL) process. Unstructured data from SPM and SP2D documents were automatically extracted with high accuracy, achieving an average of 98.52% for SPM using a combination of OCR and PDFPlumber, and 100% for SP2D extracted using PDFPlumber. The extraction results were stored in a data warehouse, then transformed using Apache Spark and loaded into data marts. The ETL process was automated using Apache Airflow, which operated reliably according to dependencies. The processed data were presented through an interactive Looker Studio dashboard in real-time, supporting efficient archive management and more informed decision-making. This study not only provides a solution to existing archival management problems but also opens opportunities for further development in the application of big data technologies and business process automation in the public sector.

Keywords: *Data Warehouse, Optical Character Recognition (OCR), Extract Transform Load (ETL), Automated Pipeline, Financial Disbursement*

ABSTRAK

Di era digital, pengelolaan data arsip menjadi tantangan bagi banyak instansi, termasuk Badan Pusat Statistik (BPS) Kabupaten Sukabumi, terutama ketika dokumen berupa PDF yang tidak terstruktur. Penelitian ini mengembangkan pipeline data dengan mengintegrasikan teknologi *Optical Character Recognition* (OCR) dan proses *Extract Transform Load* (ETL) secara efektif. Data tidak terstruktur dari dokumen SPM dan SP2D berhasil diekstraksi secara otomatis dengan tingkat akurasi yang tinggi, yaitu rata-rata 98,52% untuk SPM menggunakan kombinasi OCR dan PDFPlumber, serta 100% untuk SP2D yang diekstraksi menggunakan PDFPlumber. Hasil ekstraksi disimpan dalam *data warehouse*, kemudian ditransformasi menggunakan Apache Spark dan dimuat ke dalam *data marts*. Proses ETL diotomatisasi menggunakan Apache Airflow yang berjalan dengan stabil sesuai dependensi. Data yang telah diproses disajikan melalui dashboard interaktif Looker Studio secara *real-time*, sehingga mendukung efisiensi pengelolaan arsip serta pengambilan keputusan yang lebih informatif. Penelitian ini tidak hanya memberikan solusi terhadap permasalahan pengelolaan arsip yang ada, tetapi juga membuka peluang pengembangan lebih lanjut dalam penerapan teknologi *big data* dan otomasi proses bisnis di sektor publik.

Kata Kunci: *Data Warehouse, Optical Character Recognition (OCR), Extract Transform Load (ETL), Automated Data Pipeline, Pencairan Keuangan*

KATA PENGANTAR

Dengan penuh rasa syukur, penulis panjatkan puji dan syukur ke hadirat Allah SWT atas limpahan rahmat dan karunia-Nya. Shalawat serta salam semoga senantiasa tercurah kepada Nabi Muhammad SAW. Berkat izin dan pertolongan-Nya, penulis dapat menyelesaikan skripsi ini yang berjudul "INTEGRASI TEKNOLOGI OCR DENGAN PROSES ETL UNTUK OTOMATISASI DATA PIPELINE ARSIP PENCAIRAN KEUANGAN BPS KABUPATEN SUKABUMI" dengan baik.

Penulis menyadari bahwa skripsi ini masih jauh dari kesempurnaan, baik dalam penyajian, analisis permasalahan, maupun solusi yang ditawarkan. Oleh karena itu, penulis sangat mengharapkan kritik dan saran yang membangun demi perbaikan di masa mendatang. Semoga skripsi ini dapat memberikan manfaat, khususnya bagi penulis sendiri dan umumnya bagi siapa saja yang membutuhkan.

Dalam proses penyusunan skripsi ini, penulis mendapatkan banyak dukungan, bimbingan, serta bantuan dari berbagai pihak. Dengan demikian, dalam kesempatan ini, penulis ingin mengungkapkan rasa terima kasih yang mendalam kepada:

1. Bapak Dr. Kurniawan, S.T., M.Si., MM selaku Rektor Universitas Nusa Putra.
2. Bapak Ir. Paikun, S.T., M.T., IPM., ASEAN Eng selaku Dekan Fakultas Univeristas Nusa Putra.
3. Bapak Ir. Somantri, S.T., M.Kom selaku ketua program studi Teknik Informatika Universitas Nusa Putra.
4. Ibu Gina Purnama Insany, S.Si.T., M.Kom dan Bapak Ir. Somantri, S.T., M.Kom selaku dosen pembimbing skripsi atas segala bimbingan, arahan serta saran yang diberikan kepada penulis sehingga skripsi ini dapat diselesaikan dengan baik.
5. Bapak Alun Sujjada, S.Kom, M.T selaku pembimbing akademik, yang selalu memberikan bimbingan, arahan, dan dukungan selama proses perkuliahan.
6. Seluruh dosen/staff pengajar Fakultas Komputer Teknik dan desain program studi Teknik Informatika yang telah memberikan ilmu pengetahuan yang tak ternilai selama penulis menempuh pendidikan di Universitas Nusa Putra.

7. Bapak Muhammad Sholihin, S.E., MM selaku Kepala BPS Kabupaten Sukabumi, yang telah berkenan memberikan izin kepada penulis untuk melaksanakan penelitian ini. Penulis sangat mengapresiasi dukungan serta kerja sama yang telah diberikan, yang sangat membantu kelancaran proses penelitian dan memungkinkan penulis untuk memperoleh data yang diperlukan secara efektif.
8. Penulis juga ingin menyampaikan rasa terima kasih yang tulus kepada seluruh jajaran satuan kerja (satker), staf, serta petugas keamanan (satpam) di BPS Kabupaten Sukabumi atas segala bantuan dan dukungan yang diberikan selama penelitian ini. Secara khusus, penulis berterima kasih kepada Ibu Yuyu Yustikarini, S.Si selaku Kasubbag Umum, Ibu Ratna Rosmayanti, S.ST selaku Kepala IPDS, serta Bapak Iwan selaku pembimbing saat magang, atas bimbingan dan arahan yang sangat membantu dalam proses penelitian ini.
9. Kepada seluruh keluarga, terutama orang tua, serta teman yang selalu memberikan dukungan penuh dalam menyelesaikan skripsi ini. Semangat, dorongan, dan motivasi yang mereka berikan sangat berharga serta menjadi sumber kekuatan yang tak ternilai selama proses penulisan ini.

Sebagai penutup, penulis berharap semoga Allah SWT membalas segala kebaikan semua pihak yang telah memberikan bantuan. Semoga skripsi ini membawa manfaat bagi pengembangan ilmu.

Sukabumi, 18 Juni 2025

Penulis

Muhamamd Raihan Izharul Haq

HALAMAN PERNYATAAN PERSETUJUAN
PUBLIKASI SKRIPSI UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik UNIVERSITAS NUSA PUTRA, Saya yang bertanda tangan di bawah ini:

NAMA : MUHAMMAD RAIHAN IZHARUL HAQ

NIM 20210040075

PROGRAM STUDI : TEKNIK INFORMATIKA

Jenis karya : Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Nusa Putra **Hak Bebas Royalti Non Eksklusif (*Non-exclusive Royalty- Free Right*)** atas karya ilmiah saya yang berjudul:

“INTEGRASI TEKNOLOGI OCR DENGAN PROSES ETL UNTUK OTOMATISASI DATA PIPELINE ARSIP PENCAIRAN KEUANGAN BPS KABUPATEN SUKABUMI”

Beserta perangkat yang ada (jika diperlukan). Dengan hak bebas royalti noneksklusif ini Universitas Nusa Putra berhak menyimpan, mengalihmedia/format-kan, mengelola dalam bentuk pangkalan data (database), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik hak cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Sukabumi

Pada tanggal : 30 Juni 2025

Yang menyatakan

Muhammad Raihan Izharul Haq

DAFTAR ISI

HALAMAN SAMPUL	i
HALAMAN JUDUL.....	i
PERNYATAAN PENULIS.....	ii
PENGESAHAN SKRIPSI	iii
ABSTRACT.....	iv
ABSTRAK.....	v
KATA PENGANTAR.....	vi
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI	viii
DAFTAR ISI	ix
DAFTAR TABEL	xi
DAFTAR GAMBAR	xiii
BAB I PENDAHULUAN	1
1.1 Latar Belakang Penelitian.....	1
1.2 Rumusan Masalah	3
1.3 Batasan Masalah.....	3
1.4 Tujuan Penelitian.....	4
1.5 Manfaat Penelitian.....	5
1.5.1 Bagi Peneliti	5
1.5.2 Bagi BPS Kabupaten Sukabumi.....	5
1.5.3 Bagi Masyarakat.....	5
1.6 Sistematika Penelitian	6
BAB II TINJAUAN PUSTAKA	7
2.1 Penelitian Terkait.....	7
2.2 Landasan Teori	10
2.2.1 Badan Pusat Statistik	10
2.2.2 Arsip Pencairan Keuangan	12
2.2.3 <i>Data Engineering</i>	13
2.2.4 Proses ETL	14
2.2.5 <i>Data Warehouse</i>	14
2.2.5 <i>Optical Character Recognition (OCR)</i>	20
2.2.6 Docker	20
2.2.7 Apache Airflow.....	21
2.2.8 Python.....	23

2.2.9 PostgreSQL.....	24
2.2.10 Apache Spark.....	25
2.2.11 Looker Studio	26
2.3 Kerangka Berpikir	27
BAB III METODOLOGI PENELITIAN.....	29
3.1 Tahapan Penelitian.....	29
3.2 Metode Penelitian.....	29
3.3 Pengumpulan Data.....	32
3.3.1 Observasi	32
3.3.2 Wawancara	32
3.3.3 Studi Pustaka	33
3.4 Analisis Kebutuhan Sistem.....	33
3.5 Perencanaan dan Desain Arsitektur Data Pipeline	35
3.6 Metode OCR.....	40
3.7 Metode PDFPlumber	43
3.8 Metode Proses ETL 1	44
3.8.1 Kasus SPM	44
3.8.2 Kasus SP2D	59
3.9 Metode Proses ETL 2	68
3.10 Metode Orkestrasi <i>Data Pipeline</i>	73
3.10.1 Proses DAG SPM	74
3.10.2 Proses DAG SP2D.....	74
3.10.3 Proses DAG <i>Load to Marts</i>	75
3.11 Metode Pengujian	75
BAB IV HASIL DAN PEMBAHASAN	77
4.1 Implementasi	77
4.2 Hasil Pengujian.....	90
4.2.1 Pengujian <i>Black-box</i>	90
4.2.2 Pengujian Akurasi Model Ekstraksi Data.....	108
BAB V PENUTUP	111
5.1 Kesimpulan.....	111
5.2 Saran.....	111
DAFTAR PUSTAKA.....	113

DAFTAR TABEL

Tabel 2.1 Penelitian Terkait.....	7
Tabel 3.1 Kebutuhan Hardware.....	34
Tabel 3.2 Kebutuhan Software	34
Tabel 4.1 Pengujian Ekstraksi PDF SPM Digital Valid	90
Tabel 4.2 Pengujian Ekstraksi PDF SPM Scan Valid	91
Tabel 4.3 Pengujian Ekstraksi PDF SPM Digital Tidak Valid	92
Tabel 4.4 Pengujian Ekstraksi PDF SPM Scan Tidak Valid	93
Tabel 4.5 Pengujian SPM Koneksi ke Data Warehouse Berhasil.....	94
Tabel 4.6 Pengujian SPM Koneksi ke Data Warehouse Gagal	95
Tabel 4.7 Pengujian Validasi Field Penting SPM.....	95
Tabel 4.8 Pengujian Validasi Eksistensi Data SPM.....	95
Tabel 4.9 Pengujian Insert Data SPM Valid.....	96
Tabel 4.10 Pengujian SP2D Validasi Baris Header	96
Tabel 4.11 Pengujian Ekstraksi data SP2D Valid.....	97
Tabel 4.12 Pengujian Ekstraksi Data SP2D Tidak Valid.....	98
Tabel 4.13 Pengujian SPM Koneksi ke Data Warehouse Berhasil	98
Tabel 4.14 Pengujian SPM Koneksi ke Data Warehouse Gagal	98
Tabel 4.15 Pengujian Validasi Field Penting SP2D.....	99
Tabel 4.16 Pengujian Validasi Eksistensi data SP2D	99
Tabel 4.17 Pengujian Insert Data SP2D Valid	100
Tabel 4.18 Pengujian ETL 2 Koneksi ke Data Warehouse Berhasil.....	100
Tabel 4.19 Pengujian ETL 2 Koneksi ke Data Warehouse Gagal	100
Tabel 4.20 Pengujian membaca dan menyimpan data dalam partisi.....	101
Tabel 4.21 Pengujian ETL 2 Koneksi ke Data Marts Berhasil.....	101
Tabel 4.22 Pengujian ETL 2 Koneksi ke Data Marts Gagal.....	102
Tabel 4.23 Pengujian unggah Data ke Spreadsheet	102
Tabel 4.24 Pengujian menjalankan task DAG SPM Berhasil	103
Tabel 4.25 Pengujian Menjalankan Tasks DAG SPM Gagal	104
Tabel 4.26 Pengujian Penjadwalan DAG SPM	105
Tabel 4.27 Pengujian Menjalankan Tasks DAG SP2D Berhasil	105

Tabel 4.28 Pengujian Menjalankan Tasks DAG SP2D Gagal	106
Tabel 4.29 Pengujian Penjadwalan DAG SP2D.....	106
Tabel 4.30 Pengujian Menjalankan Tasks DAG Load to Marts Berhasil	107
Tabel 4.31 Pengujian Menjalankan Tasks DAG Load to Marts Gagal.....	108
Tabel 4.32 Evaluasi Pengujian Model Ekstraksi Data SPM.....	109



DAFTAR GAMBAR

Gambar 2.1 Kantor Badan Pusat Statistik (BPS) Kabupaten Sukabumi.....	11
Gambar 2.2 Contoh Perbedaan OLTP dengan OLAP	17
Gambar 2.3 Star Schema	18
Gambar 2.4 Snowflake Schema	19
Gambar 2.5 Logo Docker	21
Gambar 2.6 Logo Apache Airflow	22
Gambar 2.7 Contoh Proses DAG sederhana	22
Gambar 2.8 Logo Python	23
Gambar 2.9 Logo PostgreSQL	25
Gambar 2.10 Logo Apache Spark	26
Gambar 2.11 Logo Looker Studio.....	26
Gambar 2.12 Gambar Kerangka Berpikir	27
Gambar 3.1 Flowchart Tahapan Penelitian.....	29
Gambar 3.2 Konsep Pengembangan Model ADDIE.....	30
Gambar 3.3 Diagram Arsitektur Data Pipeline.....	36
Gambar 3.4 Model data warehouse.....	37
Gambar 3.5 Desain Dashboard Perbedaan Tren SP2D dan SPM.....	39
Gambar 3.6 Desain Dashboard Kategori SP2D	39
Gambar 3.7 Desain Dashboard Kategori SPM.....	40
Gambar 3.8 Proses Mesin Tesseract OCR.....	42
Gambar 3.9 Flowchart Extract and Transform SPM	45
Gambar 3.10 Deteksi kotak SPM.....	53
Gambar 3.11 Cropping 3 bagian penting SPM.....	54
Gambar 3.12 Cropping 3 bagian untuk jenis SPM tertentu.....	55
Gambar 3.13 Flow Diagram Transform and Load SPM	57
Gambar 3.14 Flowchart Extract and Transform SP2D.....	60
Gambar 3.15 Halaman dengan tabel SP2D	61
Gambar 3.16 Flow Diagram Transform and Load SP2D	66
Gambar 3.17 Flowchart Extract and Transform Pada Proses ETL 2.....	68
Gambar 3.18 Flowchart Transform and Load Proses ETL 2.....	71

Gambar 3.19 Flowchart Orkestrasi Data Pipeline	74
Gambar 3.20 Flowchart DAG SPM	74
Gambar 3.21 Flowchart DAG SP2D	75
Gambar 3.22 Flowchart DAG Load to Marts.....	75
Gambar 4.1 Arsitektur Logical Layer Data Pipeline	77
Gambar 4.2 Halaman Dokumen SPM Scan Hasil Cropping.....	80
Gambar 4.3 Hasil Ekstraksi Tesseract OCR Dokumen SPM Scan	80
Gambar 4.4 Halaman Dokumen SPM Digital.....	81
Gambar 4.5 Hasil Ekstraksi PDFPlumber Dokumen SPM Digital	81
Gambar 4.6 Halaman Dokumen SP2D Digital.....	82
Gambar 4.7 Hasil Ekstraksi PDFPlumber Dokumen SP2D Digital	82
Gambar 4.8 Entitiy Relationship Diagram (ERD) Data Warehouse	83
Gambar 4.9 Menu Containers Docker Desktop	83
Gambar 4.10 Proyek Dijalankan dengan PowerShell	84
Gambar 4.11 Halaman Utama Airflow.....	84
Gambar 4.12 Koneksi ke Data Warehouse.....	85
Gambar 4.13 Koneksi ke Data Marts.....	86
Gambar 4.14 Proses DAG SPM.....	87
Gambar 4.15 Proses DAG SP2D.....	87
Gambar 4.16 Proses DAG Load to Marts	88
Gambar 4.17 Data Marts Spreadsheets.....	89
Gambar 4.18 Dashboard Looker Studio.....	90

BAB I

PENDAHULUAN

1.1 Latar Belakang Penelitian

Pengelolaan arsip merupakan salah satu aspek penting dalam mendukung operasional suatu organisasi atau instansi. Arsip yang dikelola dengan baik tidak hanya memudahkan pencarian informasi, tetapi juga mendukung pengambilan keputusan yang lebih efektif. Arsip memiliki peran krusial dalam menjaga keberlanjutan sebuah lembaga, baik untuk operasional saat ini maupun perencanaan masa depan [1]. Pada BPS Kabupaten Sukabumi, arsip pencairan keuangan yang digunakan dalam penelitian ini meliputi Surat Perintah Membayar (SPM) dan Surat Perintah Pencairan Dana (SP2D). SPM merupakan dokumen yang berisi perintah kepada bendahara untuk melakukan pembayaran atas suatu transaksi, sedangkan SP2D adalah dokumen yang diterbitkan oleh Kantor Pelayanan Perbendaharaan Negara (KPPN) sebagai dasar pencairan dana dari kas negara ke rekening penerima [2]. Dokumen pencairan keuangan tersebut masih berbentuk dalam format yang tidak terstruktur. Proses pengolahan data dari dokumen seperti ini sering kali dilakukan secara manual, yang memakan waktu, mengurangi produktivitas, dan kurang efisien. Masalah ini menjadi semakin kompleks ketika volume data arsip yang harus diolah terus bertambah, sehingga menuntut adanya solusi yang lebih otomatis dan terstruktur.

Pengelolaan data semakin kompleks seiring dengan berkembangnya konsep big data, yang memiliki tiga karakteristik utama: kecepatan, ukuran, dan keberagaman. Variasi data ini tidak hanya mencakup perbedaan dalam jumlah dan skala, tetapi juga dalam bentuknya, yang dapat berupa data terstruktur maupun data tidak terstruktur [3]. Salah satu teknologi yang menonjol adalah *Optical Character Recognition* (OCR), yang memungkinkan ekstraksi data tidak terstruktur berupa teks dari dokumen hasil pemindaian. Dengan mengintegrasikan OCR ke dalam proses *Extract Transform Load* (ETL), sebuah sistem dapat dirancang untuk mengotomatisasi alur pengolahan data, mulai dari ekstraksi informasi, menyimpannya ke dalam format yang lebih terstruktur, hingga menyajikan data

dalam bentuk yang siap digunakan, memungkinkan pengguna akhir untuk menganalisis dan membuat keputusan secara efektif. Teknologi data pipeline ETL yang terotomasi ini tidak hanya mampu meningkatkan efisiensi proses, tetapi juga memungkinkan data untuk dianalisis secara lebih mendalam [4].

Integrasi teknologi OCR dengan data pipeline ETL menawarkan solusi yang inovatif untuk mengelola dokumen arsip tidak terstruktur. Dalam proses ini, tahap *Extract* akan memanfaatkan teknologi OCR untuk mengekstrak teks dari dokumen PDF hasil pemindaian, sementara dokumen PDF yang sudah memiliki teks dapat diekstrak langsung menggunakan library Python, yaitu PDFPlumber.

Setelah data diekstraksi, tahap *Transform* akan dilakukan menggunakan Python untuk membersihkan, memvalidasi, dan menyesuaikan format data sesuai kebutuhan. Selanjutnya, pada tahap *Load*, data yang telah diproses akan diunggah terlebih dahulu ke *data warehouse* PostgreSQL sebagai penyimpanan utama. Dengan menggunakan *data warehouse*, data history di dalamnya dapat dianalisis secara lebih mendalam dengan hasil yang akurat dan konsisten, sehingga mendukung pengambilan keputusan yang lebih efektif [5].

Kemudian, untuk keperluan analisis, data akan diekstrak dari PostgreSQL dan diproses menggunakan Apache Spark, yang bertanggung jawab untuk mengagregasi dan mengolah data sebelum dimuat ke dalam *data marts* dalam format spreadsheet. Data dalam spreadsheet ini nantinya akan digunakan sebagai sumber untuk dashboard interaktif di Looker Studio, memungkinkan pengguna untuk melakukan analisis secara visual dengan informasi yang lebih terstruktur dan mudah diakses.

Proses ETL dalam sistem ini diimplementasikan dengan pendekatan batch processing di mana data diekstraksi, ditransformasikan, dan dimuat dalam interval waktu tertentu secara terjadwal [6]. Dengan memanfaatkan Apache Airflow sebagai tool orkestrasi, sistem dapat memastikan alur kerja yang efisien serta mengelola dan menjadwalkan tugas secara otomatis dan terintegrasi.

Diharapkan dengan adanya penelitian penulis yang berjudul “Integrasi Teknologi OCR dengan Proses ETL untuk Otomatisasi Data Pipeline Arsip

Pencairan Keuangan BPS Kabupaten Sukabumi”, pengelolaan arsip dokumen keuangan yang tidak terstruktur di BPS Kabupaten Sukabumi dapat dilakukan dengan lebih cepat dan efisien. Selain itu, sistem ini dirancang untuk menyimpan histori data secara terstruktur menggunakan *data warehouse*, sehingga memungkinkan pelacakan informasi dari waktu ke waktu dan mendukung analisis data jangka panjang. Dengan adanya penyimpanan yang terpusat dan historis ini, sistem diharapkan dapat memberikan kemudahan dalam analisis data melalui dashboard interaktif yang informatif, sehingga mendukung pengambilan keputusan yang lebih tepat waktu dan berbasis data. Dengan solusi ini, instansi dapat meningkatkan produktivitas kerja serta menghadirkan layanan yang lebih baik kepada masyarakat.

1.2 Rumusan Masalah

Adapun fokus permasalahan yang dirumuskan dalam penelitian ini adalah sebagai berikut:

1. Bagaimana integrasi teknologi OCR dapat diimplementasikan dalam proses ETL untuk mengolah data tidak terstruktur dari dokumen arsip pencairan keuangan hingga menjadi data terstruktur, serta bagaimana Apache Spark digunakan dalam ETL lanjutan untuk mentransformasi data ke tahap *data marts*?
2. Bagaimana tingkat akurasi model dalam mengekstrak teks dari dokumen arsip pencairan keuangan yang tidak terstruktur?
3. Bagaimana penggunaan Apache Airflow dalam mengotomatiskan pipeline ETL untuk pengolahan data arsip pencairan keuangan?
4. Bagaimana penggunaan dashboard Looker Studio dalam visualisasi data hasil pemrosesan untuk mendukung pengambilan keputusan?

1.3 Batasan Masalah

Agar penelitian ini tetap fokus dan dapat diselesaikan sesuai waktu dan sumber daya yang tersedia, penelitian ini dibatasi pada hal-hal berikut:

1. Penelitian ini hanya mencakup integrasi proses ETL dan OCR untuk mengolah arsip pencairan keuangan dalam bentuk dokumen PDF yang tidak

dilindungi sandi (password), baik yang berisi teks maupun hasil pemindaian (scan) berupa gambar.

2. Implementasi dilakukan menggunakan arsip pencairan keuangan yang tersedia di BPS Kabupaten Sukabumi sebagai studi kasus.
3. Data yang digunakan dalam penelitian ini dibatasi pada dokumen pencairan keuangan dari bulan Januari sampai Oktober 2024, yang terdiri atas 248 dokumen SPM (242 dokumen berbasis scan dan 6 dokumen berbasis digital), serta 10 dokumen SP2D. Pengujian dilakukan pada data arsip pencairan keuangan tertentu yang relevan dengan kebutuhan data pipeline di BPS Kabupaten Sukabumi, tidak mencakup seluruh jenis arsip secara menyeluruh.
4. Pipeline ETL diimplementasikan menggunakan Apache Airflow yang dikelola dengan Astronomer (Astro) sebagai platform orkestrasi workflow, untuk memastikan alur kerja yang lebih terstruktur dan mudah dipantau.
5. Pengembangan dan eksekusi dilakukan dalam environment berbasis Docker, yang memungkinkan pengelolaan dependency dan isolasi layanan dalam container.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Merancang dan mengimplementasikan integrasi teknologi OCR dalam proses ETL untuk mengonversi data tidak terstruktur dari dokumen arsip pencairan keuangan menjadi data terstruktur, serta menganalisis peran Apache Spark dalam ETL lanjutan untuk mentransformasi data ke tahap *data marts*.
2. Menguji dan menganalisis tingkat akurasi teknologi OCR dalam mengekstrak teks dari dokumen arsip pencairan keuangan yang tidak terstruktur guna memastikan kualitas data yang dihasilkan.
3. Mengembangkan dan mengimplementasikan penggunaan Apache Airflow dalam mengotomatiskan pipeline ETL untuk pengolahan data arsip pencairan keuangan.

4. Menganalisis penerapan dashboard Looker Studio dalam menyajikan data hasil pemrosesan untuk mendukung pengambilan keputusan berbasis data secara lebih informatif dan interaktif.

1.5 Manfaat Penelitian

1.5.1 Bagi Peneliti

1. Penelitian ini memberikan kesempatan bagi peneliti untuk mengembangkan keterampilan teknis di bidang teknologi OCR, proses ETL, *data warehouse*, serta visualisasi data, yang dapat diterapkan dalam berbagai skenario pengelolaan data tidak terstruktur.
2. Peneliti dapat memperdalam wawasan tentang penerapan teknologi OCR yang terintegrasi dengan otomatisasi data pipeline dalam pengelolaan arsip dokumen tidak terstruktur, serta memahami tantangan dan solusi dalam proses tersebut.
3. Penelitian ini memberikan pengalaman langsung dalam merancang, mengembangkan, dan mengoptimalkan solusi berbasis teknologi untuk meningkatkan efisiensi pengolahan data arsip pencairan keuangan.

1.5.2 Bagi BPS Kabupaten Sukabumi

1. Sistem yang dirancang dapat meningkatkan efisiensi dan produktivitas dalam pengelolaan arsip dokumen keuangan yang tidak terstruktur.
2. Membantu BPS Kabupaten Sukabumi dalam mengurangi beban kerja manual dalam pengolahan dokumen, sehingga tenaga kerja dapat difokuskan pada tugas-tugas strategis lainnya.
3. Mendukung proses pengambilan keputusan berbasis data dengan menyediakan informasi yang lebih cepat, akurat, dan terstruktur melalui dashboard interaktif.

1.5.3 Bagi Masyarakat

1. Solusi yang dihasilkan dapat menjadi inspirasi atau model bagi institusi lain dalam mengelola arsip dokumen tidak terstruktur di sektor publik maupun swasta.

1.6 Sistematika Penelitian

Memberikan gambaran secara garis besar mengenai isi dari masing-masing bab dalam proposal skripsi ini. Sistematika penulisan laporan ini adalah sebagai berikut:

BAB I : PENDAHULUAN

Bab ini menjelaskan latar belakang yang mendasari pemilihan judul “Integrasi Teknologi OCR dengan Proses ETL untuk Otomatisasi Data Pipeline Arsip BPS Kabupaten Sukabumi”. Selain itu, bab ini mencakup perumusan masalah, ruang lingkup penelitian, tujuan yang ingin dicapai, manfaat yang diharapkan, serta sistematika penulisan laporan sebagai panduan dalam membaca penelitian ini.

BAB II : TINJAUAN PUSTAKA

Bab ini menyajikan landasan teori yang mendukung penelitian, termasuk konsep-konsep dasar serta penjelasan mengenai teknologi dan komponen yang digunakan dalam penelitian ini.

BAB III : METODOLOGI PENELITIAN

Bab ini menguraikan langkah-langkah yang dilakukan dalam penelitian, mulai dari perencanaan hingga implementasi. Selain itu, bab ini juga menjelaskan proses perancangan sistem yang dikembangkan.

BAB IV : HASIL DAN PEMBAHASAN

Bab ini membahas hasil dari penelitian yang telah dilakukan, termasuk analisis dan evaluasi terhadap sistem yang dirancang. Setiap temuan diuraikan secara rinci untuk menunjukkan efektivitas solusi yang diusulkan..

BAB V : PENUTUP

Bab terakhir ini berisi kesimpulan yang diperoleh dari penelitian, berdasarkan analisis dan implementasi yang telah dilakukan. Selain itu, bab ini juga mencantumkan saran yang dapat menjadi rekomendasi untuk pengembangan atau penelitian lanjutan di masa mendatang.

BAB V

PENUTUP

1.1 Kesimpulan

Penelitian ini berhasil mengintegrasikan teknologi OCR ke dalam proses ETL untuk mengolah dokumen arsip pencairan keuangan pada BPS Kabupaten Sukabumi, baik dalam bentuk hasil *scan* maupun dokumen digital. Data tidak terstruktur dari dokumen SPM dan SP2D berhasil diekstraksi menjadi data terstruktur, kemudian dimuat ke dalam *data warehouse* dan ditransformasi ke *data marts* menggunakan Apache Spark.

Model ekstraksi data menunjukkan tingkat akurasi yang sangat tinggi, dengan rata-rata kemiripan mencapai 98,52% untuk dokumen SPM yang diekstrak menggunakan kombinasi OCR dan PDFPlumber, serta 100% untuk dokumen SP2D yang diekstrak sepenuhnya menggunakan PDFPlumber. Tingginya akurasi pada kedua jenis dokumen tersebut membuktikan bahwa pendekatan yang digunakan andal untuk mendukung proses digitalisasi dan otomatisasi data.

Proses ETL berhasil diotomatisasi dengan Apache Airflow melalui tiga DAG utama. Seluruh proses berjalan secara terjadwal dan berurutan sesuai dengan dependensi yang telah ditetapkan, menunjukkan bahwa orkestrasi sistem berlangsung secara stabil dan andal.

Hasil akhir dari pengolahan data disajikan melalui dashboard interaktif pada Looker Studio secara *real-time*. Dashboard ini mempermudah pemantauan dan evaluasi data pencairan keuangan, serta mendukung pengambilan keputusan secara lebih efisien dan informatif.

1.2 Saran

Penelitian ini telah berhasil mengembangkan sistem otomatisasi pipeline data arsip pencairan keuangan, namun masih diperlukan peningkatan dari sisi infrastruktur dan skalabilitas untuk mendukung implementasi di lingkungan produksi berskala besar. Disarankan melakukan migrasi dari lingkungan lokal

berbasis Docker ke platform *cloud computing* seperti Amazon EC2, serta memanfaatkan layanan penyimpanan seperti Amazon S3 dan basis data terkelola seperti Amazon RDS atau Google Cloud SQL guna meningkatkan fleksibilitas, efisiensi, dan integrasi sistem.

Dengan dukungan infrastruktur yang lebih modern, sistem ini memiliki potensi untuk dikembangkan lebih lanjut dalam mendukung digitalisasi arsip dan otomatisasi birokrasi di instansi pemerintah, serta mempercepat transformasi digital di sektor publik.



DAFTAR PUSTAKA

- [1] T. Darmansah, M. N. Agung, S. S. Hasbih, and N. Lucky Tirta, “Tantangan dan Solusi dalam Pengelolaan arsip di era digital,” *J. Ekon. dan Bisnis Digit.*, vol. 02, no. 01, p. 5, 2023.
- [2] R. I. Bakari, H. Karamoy, and R. Lambey, “Analisis Prosedur Pencairan Dana Langsung (LS) Pada Kantor Pelayanan Perbendaharaan Negara (KPPN) Manado,” *J. LPPM Bid. EkoSosKum(Ekonomi, Sos. Budaya Hukum)*, vol. 5, no. 2, pp. 941–948, 2022.
- [3] Siti Sarah Sobariah Lestari, Gina Purnama Insany, Dede Sukmawan, and Faiz Dzulfikar Yusuf, “Mengatasi Permasalahan High Dimensional Space dalam Klasifikasi Multikelas Big Data pada Data Gambar dengan DCSVM,” *J. RESTIKOM Ris. Tek. Inform. dan Komput.*, vol. 5, no. 3, pp. 340–351, 2023, doi: 10.52005/restikom.v5i3.259.
- [4] Murtiwiayati, A. Hansel, and S. Leli, “Implementasi Data Warehouse dan Business Intelligence Menggunakan Pentaho dan Metabase untuk Membuat Dashboard Visualisasi Kinerja Penjualan E-Commerce Wish,” *J. Penelit. Teknol. Inf. dan Sains*, vol. Volume. 2, p. 9, 2024.
- [5] M. Adrezo and E. Ermatita, “Implementasi Pentaho Pada Perancangan Data Warehouse Perusahaan Jasa Pengiriman (PT. Tiki Palembang),” *J. Tek. Inform. dan Sist. Inf.*, vol. 10, no. 2, pp. 2407–4322, 2023, [Online]. Available: <http://jurnal.mdp.ac.id>
- [6] E. Sanchez, “What Is Batch ETL Processing? The Only Guide You Need,” Skyvia. Accessed: Mar. 30, 2025. [Online]. Available: <https://blog.skyvia.com/batch-etl-processing/>
- [7] “Sejarah Badan Pusat Statistik.” Accessed: Apr. 01, 2025. [Online]. Available: <https://ppid.bps.go.id/app/konten/0000/Profil-BPS.html>
- [8] S. Febrina and H. Syamsul, “Analisis Kinerja Badan Pusat Statistik (BPS) Dalam Menyediakan Data Statistik Yang Akurat,” *J. Ilm. Wahana Pendidik.*,

- vol. 10, no. 19, pp. 365–368, 2024, [Online]. Available: <https://jurnal.peneliti.net/index.php/JIWP>
- [9] R. Siregar, “Transformasi Digital Badan Pusat Statistik Indonesia Pada Era Big Data,” *Madani J. Ilm. Multidisiplin*, vol. 1, no. 1, pp. 581–586, 2023, [Online]. Available: <https://jurnal.penerbitdaarulhuda.my.id/index.php/MAJIM/article/view/1260%0Ahttps://jurnal.penerbitdaarulhuda.my.id/index.php/MAJIM/article/viewFile/1260/1312>
- [10] Z. Fatimatuz and R. Nur, “Analisis Sistem Pengolahan Data Berbasis Web pada Badan Pusat Statistik Kabupaten X,” *Manufaktur Publ. Sub Rumpun Ilmu Keteknikan Ind.*, vol. 2, no. 1, pp. 01–13, 2024, doi: 10.61132/manufaktur.v2i1.119.
- [11] “SP2D: Definisi, Proses, Fungsi, dan Ketentuan Penerbitan SPM.” Accessed: May 20, 2025. [Online]. Available: <https://www.labamu.co.id/article/sp2d-definisi-proses-fungsi-dan-ketentuan-penerbitan-spm#:~:text=Penyerahan berkas SPM yang sesuai,dalam pencatatan transaksi dan keuangan.>
- [12] Kamila Luthfia Yahya, Iswary Amalia, and Lies Nurhaini, “Transformasi Digitalisasi Dalam Penerbitan Surat Perintah Pencairan Dana (SP2D) Badan Pengelolaan Keuangan, Pendapatan, Dan Aset Daerah Sukoharjo: Menuju Efisiensi Dan Transparansi,” *SEWAGATI J. Pengabd. Masy. Indones.*, vol. 2, no. 4, pp. 26–36, 2023, doi: 10.56910/sewagati.v2i4.1095.
- [13] A. Razi and B. Jaswari, “Sistem Informasi Penolakan Surat Perintah Membayar (Spm) Berbasis Web Di Kantor Kppn Lhokseumawe,” *J. Teknol. Terap. Sains 4.0*, vol. 2, no. 1, p. 413, 2021, doi: 10.29103/tts.v2i1.3893.
- [14] N. Riza, M. Z. Aulia, P. B. Kolin, and K. Mustaqim, “ANALISIS FAKTOR PENGARUH TERHADAP PENGHASILAN PROFESI DATA ENGINEER MENGGUNAKAN METODE REGRESI LINEAR BERGANDA,” *JITET (Jurnal Inform. dan Tek. Elektro Ter.*, vol. 13, no. 1, p. 9, 2024.

- [15] H. Yamjala, "The Role of Data Engineering in AI and Machine Learning Projects," Dataversity. Accessed: May 19, 2025. [Online]. Available: <https://www.dataversity.net/the-role-of-data-engineering-in-ai-and-machine-learning-projects/>
- [16] R. Abdul Ghani and R. Kurniawan, "Implementasi Extract, Transform, Load Process Pada Perancangan Data Warehouse Terkait Kualitas Pendidikan Di Kabupaten Serang," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 2, pp. 2083–2090, 2024, doi: 10.36040/jati.v8i2.9081.
- [17] A. Fauzi, A. W. Noor, L. N. Ardyansyah, and J. B. Semesta, "Kajian Penerapan Arsitektur Data Warehouse dalam Bisnis Intelijen pada Pengambilan Keputusan Bisnis," *JEMSI (Jurnal Ekon. Manaj. Sist. Informasi)*, vol. 4, no. 5, pp. 868–875, 2023, [Online]. Available: <https://www.dinastirev.org/JEMSI/article/download/1501/936>
- [18] R. Atsila Imanda *et al.*, "Pengaruh Data Warehouse Terhadap Pengambilan Keputusan," *J. Portofolio J. Manaj. dan Bisnis*, vol. 3, no. 1, pp. 31–39, 2024, [Online]. Available: <https://www.jurnalprisanicendekia.com/index.php/portofolio/article/view/282>
- [19] N. Adila and Andri, "Desain Dan Implementasi Data Warehouse Pada Perpustakaan Daerah Provinsi Sumatera Selatan," *J. Nas. Ilmu Komput.*, vol. 2, no. 1, pp. 33–50, 2021, doi: 10.47747/jurnalnrik.v2i1.520.
- [20] A. F. Karami, "Perancangan Arsitektur Data Warehouse Pada Industri Perkebunan Kelapa Sawit," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 2, pp. 973–983, 2022, doi: 10.35957/jatisi.v9i2.1834.
- [21] A. Khaerul and Lasimin, "BUSINESS INTELLIGENCE PADA SISTEM MANAJEMEN ASET UNUGHA DENGAN METODE ONLINE ANALYTICAL PROCESSING (OLAP)," *J. Inov. Glob.*, vol. 2, no. 10, 2024.
- [22] "Difference Between OLAP and OLTP in Databases," GeeksforGeeks. Accessed: May 19, 2025. [Online]. Available:

<https://www.geeksforgeeks.org/difference-between-olap-and-oltp-in-dbms/>

- [23] M. Hidayat, “Perancangan Dashboard Untuk Visualisasi Data Jumlah Penumpang Transjakarta,” *Insantek*, vol. 4, no. 1, pp. 32–36, 2023, doi: 10.31294/insantek.v4i1.2222.
- [24] R. R. Berahim and M. Iqbal, “Analysis and Design Optimize Data for the Depok Center Information System’s Health Services on the E-Government Data Warehouse Application using an Olap Pivot Table Analysis and Design Optimize Data for the Depok Center Information System’s Health Service,” *Int. Res. J. Adv. Eng. Sci.*, vol. 6, no. 3, pp. 218–226, 2021.
- [25] A. Undru, “Star Schema vs Snowflake Schema: 6 Key differences,” *thoughtspot.com*. Accessed: May 27, 2025. [Online]. Available: <https://www.thoughtspot.com/data-trends/data-modeling/star-schema-vs-snowflake-schema>
- [26] K. Banu, D. Andreas, W. Anggoro, and A. Setiawan, “OCR: Masa Depan Pengenalan Karakter Optik dan Dampaknya pada Kehidupan Modern,” *J. Teknol. Inf.*, vol. 9, no. 2, pp. 147–156, 2023, doi: 10.52643/jti.v9i2.3798.
- [27] Z. M. Subekti, K. Mukiman, Subandri, M. L. Sulthon, A. Sulistiyono, and R. E. Putra, “RANCANG BANGUN INFRASTRUKTUR WEB SERVER,” *J. TRIDI*, vol. 2, no. 1, pp. 144–151, 2024.
- [28] S. Dwiyatno, E. Rachmat, A. P. Sari, and O. Gustiawan, “Implementasi Virtualisasi Server Berbasis Docker Container,” *PROSISKO J. Pengemb. Ris. dan Obs. Sist. Komput.*, vol. 7, no. 2, pp. 165–175, 2020, doi: 10.30656/prosisko.v7i2.2520.
- [29] D. R. Saputra, W. Purnomo, N. Y. Setiawan, and P. Korespondensi, “Sistem Propagasi Anotasi pada Metadata Lineage untuk Manajemen Data WarehouseAnnotation Propagation System on Lineage Metadata for Data Warehouse Management,” vol. 9, no. 7, pp. 1741–1746, 2022, doi: 10.25126/jtiik.202296833.
- [30] E. E. Wahyudi *et al.*, “Akuisisi Data Prediksi Curah Hujan Secara Periodik

- Menggunakan Apache Airflow,” *J. Informatics, Inf. Syst. Softw. Eng. Appl.*, vol. 4, no. 2, pp. 1–12, 2022, doi: 10.20895/inista.v4i2.574.
- [31] B. Roepke, “Getting Started with Astronomer Airflow The Data Engineering Workhorse,” data knows all. Accessed: May 23, 2025. [Online]. Available: <https://dataknowsall.com/blog/astrointro.html>
- [32] N. N. Pandika Pinata, I. M. Sukarsa, and N. K. Dwi Rusjayanthi, “Prediksi Kecelakaan Lalu Lintas di Bali dengan XGBoost pada Python,” *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, vol. 8, no. 3, p. 188, 2020, doi: 10.24843/jim.2020.v08.i03.p04.
- [33] Adawiyah Ritonga and Yahfizham Yahfizham, “Studi Literatur Perbandingan Bahasa Pemrograman C++ dan Bahasa Pemrograman Python pada Algoritma Pemrograman,” *J. Tek. Inform. dan Teknol. Inf.*, vol. 3, no. 3, pp. 56–63, 2023, doi: 10.55606/jutiti.v3i3.2863.
- [34] S. H. Ahsa, M. B. Syahputra, A. F. F. M. Putri, and A. A. Prasetyo, “ANALISIS PERBANDINGAN PERFORMA ANTARA MySQL dan PostgreSQL,” no. September, pp. 6–7, 2023.
- [35] “Apa itu Apache Spark?” Accessed: May 10, 2025. [Online]. Available: <https://www.ibm.com/id-id/topics/apache-spark>
- [36] C. Wibawa, S. Wirawan, M. Mustikasari, and D. T. Anggraeni, “Komparasi Kecepatan Hadoop Mapreduce Dan Apache Spark Dalam Mengolah Data Teks,” *J. Ilm. Matrik*, vol. 24, no. 1, pp. 10–20, 2022, doi: 10.33557/jurnalmatrik.v24i1.1649.
- [37] T. Imam Muarif and R. Danar Dana, “Implementasi Tableau Untuk Pengembangan Visualisasi Data Pada Aplikasi Open Data Di Diskominfo Kabupaten Cirebon,” *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 8, no. 1, pp. 902–909, 2024, doi: 10.36040/jati.v8i1.8868.
- [38] A. Hadi Wijaya and W. Fitri, “Dasboard dan Visualisasi Reservasi Buka Pusa di Hotel XYZ Menggunakan Looker Studio & Google Form,” *J. Informatics Busisnes*, vol. 01, no. 04, pp. 354–359, 2024.

- [39] H. Syuhada, S. Hidayat, S. Mulyati, and A. Giri Persada, “Pengembangan Gamifikasi Pada Pelajaran Matematika Sd Dengan Metode Addie Untuk Meningkatkan Minat Belajar Siswa,” *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 9, no. 1, pp. 1–14, 2023, doi: 10.36341/rabit.v9i1.466.
- [40] Kamdan, Somantri, M. G. Sundayana, and I. L. Kharisma, “Rancang Bangun Layanan Private cloud Berbasis Infrastructure as a Service Menggunakan OpenStack dengan Metode Network Development Life Cycle(NDLC),” *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 1, pp. 252–262, 2023, doi: 10.30865/klik.v4i1.1001.
- [41] B. Suriansyah, L. F. Mz, A. I. Rachman, and G. Pratiwi, “Rekontruksi Arsitektur DataBase untuk Peningkatan Proses Load Data,” *J. MEDIA Inform. [JUMIN]*, vol. 6, no. 2, pp. 1455–1460, 2025.
- [42] M. N. Darpito, K. Firdausy, and A. Fadlil, “Perbandingan Unjuk Kerja Library Optical Character Recognition (OCR) dalam Pengenalan Teks pada Dokumen Digital,” *JIP (Jurnal Inform. Polinema)*, vol. 11, no. 3, pp. 273–282, 2025.
- [43] C. Writes, “Text Detection and Recognition with OCR and OpenCV,” Medium. Accessed: Jun. 22, 2025. [Online]. Available: <https://carl-writes.medium.com/text-detection-and-recognition-with-ocr-and-opencv-dc8e5e164d78>
- [44] I. M. D. R. Mudiarta *et al.*, “Balinese character recognition on mobile application based on tesseract open source OCR engine,” *J. Phys. Conf. Ser.*, vol. 1516, no. 1, 2020, doi: 10.1088/1742-6596/1516/1/012017.
- [45] A. E. Budiman and A. Widjaja, “Analisis Pengaruh Teks Preprocessing Terhadap Deteksi Plagiarisme Pada Dokumen Tugas Akhir,” *J. Tek. Inform. dan Sist. Inf.*, vol. 6, no. 3, pp. 475–488, 2020, doi: 10.28932/jutisi.v6i3.2892.
- [46] “pdfplumber 0.11.7,” pypi.org. Accessed: Apr. 02, 2025. [Online]. Available: <https://pypi.org/project/pdfplumber/>

- [47] F. Yusuf Alfiansyah and D. Arisandi, “Perancangan Dashboard Monitoring Status Gizi Balita di Puskesmas Sukanagalih,” *J. Ilm. Tek. Inform. dan Sist. Inf.*, pp. 1–11, 2023.

