Extracting and Processing Tabular Data from PDF to JSON

Overview of the exercise

This project involved the complex task of processing extensive PDF documents, often exceeding 2000 pages, to extract and structure critical tabular data. The raw data, rich in geographic information (GIS coordinates), streetlight details, and other metadata, was essential for streetlight-based mapping and detection systems.

Example of raw data:

Sl. no.	RR / Unique Identification number of energy meter	GIS co-ordinate of energy meter / switching point		Pole No. (BESCOM / ULB / Unique Identification no.)	GIS co-ordinates of pole (Latitude, Longitude)		Type of streetlight/ flood light	Total no. of Light In a pole	Road Category	Rated wattage	Rated wattage of proposed LED streetlight/ flood light.
		Latitude	Longitude		Latitude	Longitude					
1	BBMP/DSR/16/S1	13.0549696	77.53472742	BBMP/DSR/16/3193	13.0549696	77.53472742	MH	6	A2	400W	90
1	BBMP/DSR/16/S10	13.05796367	77.53272303	BBMP/DSR/16/67	13.05796367	77.53272303	SV	1	B2	250W	40
2		13.05803586	77.53285578	BBMP/DSR/16/68	13.05803586	77.53285578	SV	1	B2	250W	40
3		13.05808289	77.53296442	BBMP/DSR/16/69	13.05808289	77.53296442	SV	1	B2	250W	40
4		13.05806886	77.53292083	BBMP/DSR/16/70	13.05806886	77.53292083	SV	1	B2	250W	40
5		13.05818153	77.53335703	BBMP/DSR/16/118	13.05818153	77.53335703	SV	1	B1	250W	65/70
6		13.05850747	77.53330506	BBMP/DSR/16/119	13.05850747	77.53330506	SV	1	B1	250W	65/70
7		13.05848103	77.53331444	BBMP/DSR/16/120	13.05848103	77.53331444	SV	1	B1	250W	65/70
8		13.05843825	77.53362861	BBMP/DSR/16/131	13.05843825	77.53362861	LED	1	A2	90W	90
9		13.05829356	77.53367017	BBMP/DSR/16/132	13.05829356	77.53367017	LED	1	A2	90W	90
10		13.05824392	77.53360714	BBMP/DSR/16/133	13.05824392	77.53360714	LED	1	B1	90W	65/70
11		13.05846372	77.53361083	BBMP/DSR/16/134	13.05846372	77.53361083	LED	1	A2	90W	90
12		13.05822106	77.53354111	BBMP/DSR/16/135	13.05822106	77.53354111	LED	1	B1	90W	65/70
13		13.0581835	77.53341672	BBMP/DSR/16/136	13.0581835	77.53341672	LED	1	B1	90W	65/70
14		13.05812797	77.53311597	BBMP/DSR/16/137	13.05812797	77.53311597	SV	1	B1	250W	65/70
15		13.05826972	77.53308244	BBMP/DSR/16/138	13.05826972	77.53308244	SV	1	B1	250W	65/70
16		13.05843758	77.53305428	BBMP/DSR/16/139	13.05843758	77.53305428	SV	1	B1	250W	65/70
17		13.05826186	77.53243569	BBMP/DSR/16/263	13.05826186	77.53243569	SV	1	B1	250W	65/70
18		13.05882756	77.53435247	BBMP/DSR/16/1747	13.05882756	77.53435247	SV	1	B1	250W	65/70
19		13.05846797	77.53373992	BBMP/DSR/16/1748	13.05846797	77.53373992	SV	1	B1	250W	65/70
20		13.058547	77.53411778	BBMP/DSR/16/1749	13.058547	77.53411778	SV	1	B1	250W	65/70
21		13.05860317	77.53433067	BBMP/DSR/16/1750	13.05860317	77.53433067	SV	1	B1	250W	65/70
22		13.05863617	77.53443797	BBMP/DSR/16/1751	13.05863617	77.53443797	SV	1	B1	250W	65/70
23		13.05807308	77.53252986	BBMP/DSR/16/2629	13.05807308	77.53252986	SV	1	B2	250W	40
24		13.05813647	77.53272642	BBMP/DSR/16/2630	13.05813647	77.53272642	LED	1	B2	90W	40
25		13.05835317	77.5333445	BBMP/DSR/16/2632	13.05835317	77.5333445	SV	1	B1	250W	65/70
26		13.05830539	77.53271433	BBMP/DSR/16/2635	13.05830539	77.53271433	LED	1	B1	90W	65/70
27		13.05824558	77.53406908	BBMP/DSR/16/3324	13.05824558	77.53406908	LED	1	B1	24W	65/70
28		13.05848450	77.53387219	BBMP/DSR/16/3325	13.0584845	77.53387219	LED	1	B1	24W	65/70
29		13.05851153	77.53400750	BBMP/DSR/16/3326	13.05851153	77.5340075	LED	1	B1	24W	65/70
30		13.05815697	77.53371961	BBMP/DSR/16/3327	13.05815697	77.53371961	LED	1	B1	30W	65/70
1	BBMP/DSR/16/S100	13.06519106	77.53061178	BBMP/DSR/16/1334	13.06519106	77.53061178	LED	1	B1	90W	65/70
2		13.06498594	77.53064833	BBMP/DSR/16/1335	13.06498594	77.53064833	LED	1	B1	90W	65/70
3		13.06493336	77.53064533	BBMP/DSR/16/1336	13.06493336	77.53064533	LED	1	B1	90W	65/70
4		13.06503592	77.53053567	BBMP/DSR/16/1337	13.06503592	77.53053567	SV	1	B1	250W	65/70
5		13.06502058	77.53044483	BBMP/DSR/16/1338	13.06502058	77.53044483	SV	1	B2	250W	40
6		13.06503592	77.53030972	BBMP/DSR/16/1339	13.06503592	77.53030972	SV	1	B2	250W	40
7		13.06503428	77.53014408	BBMP/DSR/16/1340	13.06503428	77.53014408	SV	1	B2	250W	40

Steps Taken and the objectives:

Step 1: Split Large PDFs

1. **Optimized PDF Parsing**
   Utilized efficient libraries like PyMuPDF to extract tabular data from PDFs containing 1,000–2,000 pages.
   Optimized parsing algorithms to minimize memory usage and computational overhead.
2. **Parallel Processing**
   Implemented parallel processing with libraries like multiprocessing and Dask to distribute workloads across CPU cores.

Large PDFs were split into mostly 100 or 200 pages (mostly 250) to avoid errors and system crashes.

3. **Handling MuPDF Limitations**
   To address MuPDF's limitation in processing extremely large files, you can integrate a pre-splitting mechanism. Start by dividing the large file into smaller, manageable chunks of 100 pages or 200 pages before feeding them into MuPDF. This approach ensures smoother execution and prevents the tool from being overwhelmed by file size constraints. Various file-splitting libraries or scripts can be utilized to achieve this efficiently, depending on your specific workflow requirements.

4. **Error Mitigation and Resource Optimization**
   Added lazy loading and on-demand extraction to minimize memory usage.
   Incorporated error-handling and memory profiling to prevent performance bottleneck.
5. **Technical Stack**
   Libraries: PyMuPDF, MYPDF.io, PyPDF2,
   Techniques: Parallel processing, file splitting, lazy loading

Step 2: Clean and Format Data

1. **Load Raw Data**
   - Imported extracted data from split PDFs into CSV and JSON formats for further processing.
2. **Data Cleaning and Normalization**
   - Robust Data Cleaning: Implemented algorithms to handle null values, inconsistent formats, and unnecessary whitespace, ensuring data integrity.
   - Standardized Data Format: Transformed the cleaned data into a consistent format, enabling seamless downstream integration.
3. **Why CSV for Initial Processing**?
   - Simplicity and Compatibility: CSV files are widely supported and easy to load into data processing tools like pandas.
   - Tabular Structure: Ideal for organizing extracted data into rows and columns for quick validation and cleaning.
   - Lightweight: CSV files are compact and efficient for handling structured data without complex nesting.
4. **Why JSON for GeoJSON Conversion?**
   - Hierarchical Data Representation: JSON's nested structure aligns perfectly with GeoJSON, which requires coordinate and feature properties in a structured format.
   - Direct Integration: JSON allows for seamless conversion to GeoJSON with latitude and longitude fields directly mapped to geographic features.
   - Geospatial Compatibility: JSON facilitates the creation of GIS-compatible outputs, enabling use with mapping libraries like Leaflet or tools like QGIS.
5. **Technical Stack**
   - Libraries: pandas ,json, geojson.io

- Techniques: Data cleaning, normalization, CSV-to-JSON conversion, GeoJSON formatting

Step 3:Combining and Formatting GIS Coordinates

1 **Coordinate Extraction and Validation**

- Extracted latitude and longitude values, often spread across multiple fields, ensuring their accuracy and consistency.
- Implemented validation checks to confirm the correctness of coordinates, detecting and handling errors like missing or out-of-range values.

2 **Coordinate Reformatting**

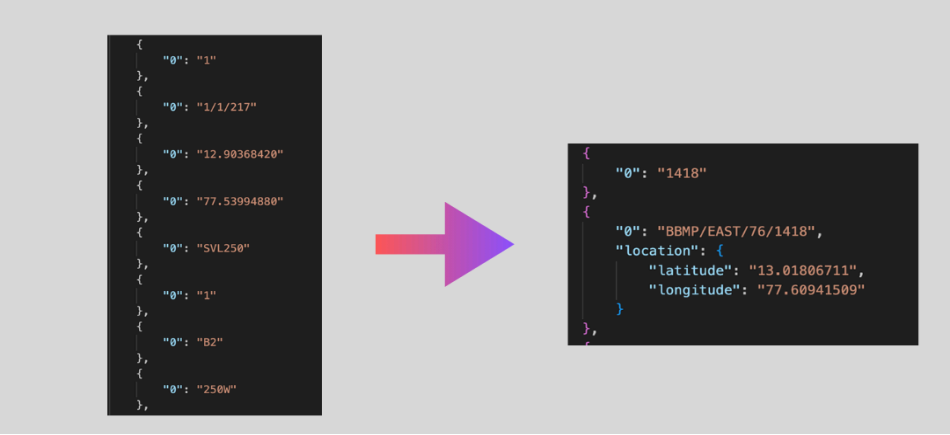
- Reformatted extracted coordinates into a standardized format (e.g., `latitude, longitude`) suitable for GIS mapping applications.
- Addressed inconsistencies caused by varying data sources or formats to ensure compatibility.

3. **Merging Coordinate Fields**

- Used a Python script to combine latitude and longitude fields into a single coordinate pair for streamlined processing.
- Libraries like pandas and geopandas were utilized for efficient merging and formatting operations.

4. **Technical Stack**

- Libraries: pandas, geopandas
- Techniques: Coordinate extraction, validation, field merging, and standardization
- the latitude and longitude values into a single coordinate pair (e.g., "latitude, longitude").



Step 4: Chunk Rows and Structure JSON

1. **Iterate Over Rows**

- Processed each row of the cleaned and formatted data individually to ensure accurate extraction and transformation.

2. **Create JSON Objects**
   - Generated JSON objects for each row by mapping column headers to relevant key-value pairs,

```
"ID": "1127",
"Pole No. (BESCOM / ULB / Unique Identification no.)": "BBMP/EAST/123/1127",
"GIS co-ordinates of pole (Latitude, Longitude)": {
   "latitude": "12.97854048",
   "longitude": "LED"
},
"Road Category": "B2",
"Rated wattage": "90W",
"Light Count": "1"
```

   like these for example .
   - ensuring clear and logical representation of the data fields.
   - Carefully handled data types (e.g., integers, floats, strings) and ensured proper formatting for consistency.
3. **Structure JSON Data**
   - Organized individual JSON objects into a structured list or array to represent the entire dataset, enabling efficient storage and retrieval.
4. **JSON Schema Design**
   - Developed a robust JSON schema to standardize the representation of extracted data, ensuring:
   - Data consistency across all records.
   - Interoperability with downstream applications such as APIs and GIS tools.
5. **Data Transformation**
   - Transformed tabular data into JSON format by mapping rows and column headers to JSON key-value pairs using libraries like pandas and json.
   - Addressed potential formatting issues (e.g., null values, nested structures) to create clean, well-structured JSON objects.
6 **Technical Stack**
   - Libraries: pandas, json

```
{
    "Type of streetlight/ flood light": "SVL250",
    "Light Count": "1",
    "Road Category": "B2",
    "Rated wattage": "40W",
    "Pole Identification Number": "19/10/217",
    "GIS co-ordinates of pole (Latitude, Longitude)": {
        "latitude": "12.91133220",
        "longitude": "77.55179660"
    }
},
```


Flowchart

Additional Considerations:

- **Error Handling**: Implement error handling mechanisms to gracefully handle potential exceptions during the processing steps.
- **Data Validation**: Consider adding data validation checks to ensure data quality and consistency.
- **Parallel Processing**: For large datasets, explore parallel processing techniques to speed up the processing time.
- **Visualization**: Use visualization tools to explore and understand the cleaned and structured data.

FINAL RESULT :

TABLE-09

Ward No 66

RR / Unique Identification number of energy meter	GIS co-ordinate of energy meter / switching point		Pole No. (BESCOM / ULB / Unique Identification no.)	GIS co-ordinates of pole (Latitude, Longitude)		Type of streetlig ht/ flood light	Road Category	Rated wattage	Total No. of Light	Rated watt proposed streetlight/ light.
	Latitude	Longitude		Latitude	Longitude					
S1	13.02809	77.57353418	BBMP/EAST/66/1	13.02848	77.57145949	LED	B1	90W	2	
			BBMP/EAST/66/2	13.02846	77.57157248	LED	A1	90W	2	
			BBMP/EAST/66/3	13.02842	77.57174179	LED	A1	90W	2	
			BBMP/EAST/66/4	13.0284	77.57188831	LED	A1	90W	2	
			BBMP/EAST/66/5	13.02836	77.57205058	LED	A1	90W	2	
			BBMP/EAST/66/6	13.02834	77.57218301	LED	A1	90W	2	
			BBMP/EAST/66/7	13.0283	77.57234562	LED	A1	90W	2	
			BBMP/EAST/66/8	13.02826	77.57248141	LED	A1	90W	2	
			BBMP/EAST/66/9	13.02824	77.57262457	LED	A1	90W	2	
			BBMP/EAST/66/10	13.02821	77.57279892	LED	A1	90W	2	
			BBMP/EAST/66/11	13.02818	77.57295951	LED	A1	90W	2	
			BBMP/EAST/66/12	13.02816	77.57308658	LED	A1	90W	2	
			BBMP/EAST/66/13	13.02812	77.57325992	LED	A1	90W	2	
			BBMP/EAST/66/14	13.02806	77.57352162	LED	A1	90W	2	
			BBMP/EAST/66/15	13.02803	77.57369913	LED	A1	90W	2	
			BBMP/EAST/66/16	13.02799	77.57385872	LED	A1	90W	2	
			BBMP/EAST/66/17	13.02798	77.57394824	LED	A1	90W	2	
			BBMP/EAST/66/18	13.02797	77.57407766	LED	A1	90W	2	
			BBMP/EAST/66/19	13.02794	77.57421613	LED	A1	90W	2	
			BBMP/EAST/66/20	13.0279	77.57447932	LED	A1	90W	2	
			BBMP/EAST/66/21	13.02786	77.57468887	LED	A1	90W	2	
S10	13.02908	77.57617783	BBMP/EAST/66/200	13.02919	77.57618252	SV	B1	250W	1	65/7(
			BBMP/EAST/66/202	13.02934	77.57618219	SV	B1	250W	1	65/7(
			BBMP/EAST/66/205	13.02948	77.57619124	SV	B1	250W	1	65/7(
			BBMP/EAST/66/208	13.0291	77.57630557	SV	A2	250W	1	90
			BBMP/EAST/66/210	13.02923	77.5763059	LED	B1	90W	1	
			BBMP/EAST/66/212	13.02943	77.57634647	SV	B1	250W	1	65/7(
			BBMP/EAST/66/215	13.02907	77.57646348	SV	A2	250W	1	90
S100	13.04054	77.56958663	BBMP/EAST/66/1875	13.04054	77.56957926	SV	B1	250W	1	65/7(
			BBMP/EAST/66/1876	13.04076	77.56960876	SV	B1	250W	1	65/7(
			BBMP/EAST/66/1877	13.04107	77.56962821	SV	B1	250W	1	65/7(
			BBMP/EAST/66/1878	13.0414	77.56962294	SV	B1	250W	1	65/7(
			BBMP/EAST/66/1879	13.04158	77.56938669	SV	B1	250W	1	65/7(
			BBMP/EAST/66/1880	13.0416	77.56944682	SV	B1	250W	1	65/7(
			BBMP/EAST/66/1881	13.04166	77.56922219	SV	B1	250W	1	65/7(
			BBMP/EAST/66/1882	13.04172	77.56894391	SV	B1	250W	1	65/7(
			BBMP/EAST/66/1883	13.04177	77.56876051	SV	B1	250W	1	65/7(
			BBMP/EAST/66/1884	13.04184	77.56815785	SV	B1	250W	1	65/7(
			BBMP/EAST/66/1885	13.04159	77.5684363	SV	B1	250W	1	65/7(
			BBMP/EAST/66/1886	13.04131	77.56848659	SV	B1	250W	1	65/7(
			BBMP/EAST/66/1887	13.0413	77.56868575	SV	B1	250W	1	65/7(
			BBMP/EAST/66/1888	13.04122	77.56894626	LED	B1	90W	1	
			BBMP/EAST/66/1889	13.04114	77.56922957	SV	B1	250W	1	65/7(
			BBMP/EAST/66/1890	13.04106	77.56952628	SV	B1	250W	1	65/7(