

1. File Check Module

This module verifies that the files received daily are as expected before processing them. It checks if the files are new, non-empty, and have a .csv extension.

```
import os

import glob

def file_check_module(directory_path, processed_files_directory):

    # List all CSV files in the directory

    all_files = glob.glob(os.path.join(directory_path, "*.csv"))

    valid_files = []

    invalid_files = []

    for file in all_files:

        # Check if the file is new

        file_name = os.path.basename(file)

        if os.path.exists(os.path.join(processed_files_directory, file_name)):

            invalid_files.append(file)

            continue

        # Check if the file is non-empty

        if os.path.getsize(file) == 0:

            invalid_files.append(file)

            continue

        # If all checks pass, add to valid files

        valid_files.append(file)

    return valid_files, invalid_files

# Example usage:
```

```
directory_path = '/path/to/daily/files'
processed_files_directory = '/path/to/processed/files'
valid_files, invalid_files = file_check_module(directory_path, processed_files_directory)

print("Valid files:", valid_files)
print("Invalid files:", invalid_files)
```

2. Data Quality Check Module

This module ensures that the data inside the files is of acceptable quality before processing. It validates phone numbers, checks for null values in specific fields, and cleans descriptive fields.

```
import pandas as pd
import re

def clean_phone_number(phone):
    phone = re.sub(r'\D', '', phone)
    if len(phone) == 10:
        return phone
    return None

def data_quality_check(file_path):
    df = pd.read_csv(file_path)

    # Validate phone numbers
    df['phone'] = df['phone'].astype(str).apply(clean_phone_number)

    # Check for null values in mandatory fields
    mandatory_fields = ['name', 'phone', 'location']
    null_rows = df[df[mandatory_fields].isnull().any(axis=1)]
    clean_rows = df.dropna(subset=mandatory_fields)
```

```
# Clean descriptive fields

df['address'] = df['address'].astype(str).apply(lambda x: re.sub(r'^a-zA-Z0-9\s,]', ', ', x))

return clean_rows, null_rows


# Example usage:

file_path = 'path/to/data_file.csv'

clean_data, bad_data = data_quality_check(file_path)

print("Clean data:\n", clean_data)
print("Bad data:\n", bad_data)


# Save clean and bad data to files

clean_data.to_csv('path/to/output/clean_data.out', index=False)
bad_data.to_csv('path/to/output/bad_data.bad', index=False)
```

The `file_check_module` function verifies if files are new, non-empty, and have a .csv extension. The `data_quality_check` function cleans phone numbers, checks for null values in mandatory fields, and cleans descriptive fields like addresses.