Report\_genereator.py

import pandas as pd

import numpy as np

import datacompy

from ydata\_profiling import ProfileReport, compare

import xlsxwriter

from datetime import datetime

import os

import zipfile

from io import BytesIO

import logging

from typing import Dict, List, Optional, Tuple

logger = logging.getLogger(\_\_name\_\_)

def generate\_datacompy\_report(source\_df: pd.DataFrame, target\_df: pd.DataFrame,

                            join\_columns: List[str], mapping\_df: pd.DataFrame,

                            join\_mappings: Dict[str, str]) -> Tuple[BytesIO, BytesIO]:

    """

    Generate a DataCompy comparison report.

    Args:

        source\_df: Source DataFrame

        target\_df: Target DataFrame

        join\_columns: List of source columns to join on

        mapping\_df: DataFrame containing column mapping information

        join\_mappings: Dictionary mapping source join columns to target join columns

    Returns:

        BytesIO object containing the report

    """

    try:

        # Get excluded columns

        excluded\_columns = mapping\_df[mapping\_df['Exclude from Comparison']]['Source Column'].tolist()

        # Create mapping dictionary from mapping\_df

        column\_mapping = dict(zip(

            mapping\_df['Source Column'],

            mapping\_df['Target Column']

        ))

        # Filter out unmapped and excluded columns

        valid\_columns = {

            src: tgt for src, tgt in column\_mapping.items()

            if tgt and not pd.isna(tgt) and src not in excluded\_columns

        }

        # Prepare DataFrames for comparison

        source\_cols = list(valid\_columns.keys())

        target\_cols = [valid\_columns[src] for src in source\_cols]

        source\_compare = source\_df[source\_cols].copy()

        target\_compare = target\_df[target\_cols].copy()

        # Rename target columns to match source columns for comparison

        target\_compare.columns = source\_cols

        # Get the target join column names

        target\_join\_columns = [join\_mappings[src] for src in join\_columns]

        # Rename join columns in target DataFrame to match source

        join\_column\_mapping = dict(zip(target\_join\_columns, join\_columns))

        target\_compare.rename(columns=join\_column\_mapping, inplace=True)

        # Ensure datetime columns are properly converted

        for col in source\_compare.columns:

            if source\_compare[col].dtype == 'datetime64[ns]':

                target\_compare[col] = pd.to\_datetime(target\_compare[col], errors='coerce')

            elif target\_compare[col].dtype == 'datetime64[ns]':

                source\_compare[col] = pd.to\_datetime(source\_compare[col], errors='coerce')

        # Ensure all columns are strings to prevent type comparison issues

        source\_compare.columns = source\_compare.columns.astype(str)

        target\_compare.columns = target\_compare.columns.astype(str)

        # Create comparison object

        try:

            comparison = datacompy.Compare(

                df1=source\_compare,

                df2=target\_compare,

                join\_columns=join\_columns,

                df1\_name='Source',

                df2\_name='Target',

                on\_index=False

            )

        except Exception as e:

            logger.error(f"Error creating comparison object: {str(e)}")

            # Try alternative comparison with converted data types

            source\_compare = source\_compare.astype(str)

            target\_compare = target\_compare.astype(str)

            comparison = datacompy.Compare(

                df1=source\_compare,

                df2=target\_compare,

                join\_columns=join\_columns,

                df1\_name='Source',

                df2\_name='Target',

                on\_index=False

            )

        # Generate report

        output = BytesIO()

        with pd.ExcelWriter(output, engine='xlsxwriter') as writer:

            # Write summary

            # Create summary DataFrame

            summary\_data = {

                'Metric': [

                    'Rows in Source',

                    'Rows in Target',

                    'Rows in Common',

                    'Rows Only in Source',

                    'Rows Only in Target',

                    'Columns Match',

                    'All Row Values Match'

                ],

                'Value': [

                    len(source\_compare),

                    len(target\_compare),

                    comparison.intersect\_rows,

                    len(comparison.df1\_unq\_rows) if hasattr(comparison, 'df1\_unq\_rows') else 0,

                    len(comparison.df2\_unq\_rows) if hasattr(comparison, 'df2\_unq\_rows') else 0,

                    comparison.all\_columns\_match(),

                    comparison.matches()

                ]

            }

            pd.DataFrame(summary\_data).to\_excel(writer, sheet\_name='Summary', index=False)

            # Write column stats

            if hasattr(comparison, 'column\_stats') and comparison.column\_stats is not None:

                if isinstance(comparison.column\_stats, pd.DataFrame):

                    comparison.column\_stats.to\_excel(writer, sheet\_name='Column Stats', index=True)

                else:

                    pd.DataFrame(comparison.column\_stats).to\_excel(writer, sheet\_name='Column Stats', index=True)

            # Write sample mismatches

            try:

                # Get mismatched rows

                df1\_unq = comparison.df1\_unq\_rows if hasattr(comparison, 'df1\_unq\_rows') else pd.DataFrame()

                df2\_unq = comparison.df2\_unq\_rows if hasattr(comparison, 'df2\_unq\_rows') else pd.DataFrame()

                # Create sample mismatches DataFrame

                if not df1\_unq.empty or not df2\_unq.empty:

                    mismatches = pd.concat([

                        df1\_unq.assign(Source='Source Only').head(5),

                        df2\_unq.assign(Source='Target Only').head(5)

                    ], ignore\_index=True)

                    mismatches.to\_excel(writer, sheet\_name='Sample Mismatches', index=True)

                else:

                    pd.DataFrame({'Message': ['No mismatches found']}).to\_excel(

                        writer, sheet\_name='Sample Mismatches', index=False)

            except Exception as e:

                logger.warning(f"Error generating sample mismatches: {str(e)}")

                pd.DataFrame({'Message': [f'Error generating sample mismatches: {str(e)}']}).to\_excel(

                    writer, sheet\_name='Sample Mismatches', index=False)

        # Generate HTML report

        html\_output = BytesIO()

        html\_report = f"""

        <html>

        <head>

            <title>DataCompy Comparison Report</title>

            <style>

                body {{ font-family: Arial, sans-serif; margin: 20px; }}

                table {{ border-collapse: collapse; width: 100%; margin-bottom: 20px; }}

                th, td {{ border: 1px solid #ddd; padding: 8px; text-align: left; }}

                th {{ background-color: #f2f2f2; }}

                .pass {{ color: green; }}

                .fail {{ color: red; }}

                .section {{ margin-bottom: 30px; }}

            </style>

        </head>

        <body>

            <h1>DataCompy Comparison Report</h1>

            <div class="section">

                <h2>Summary</h2>

                <p>Source rows: {len(source\_compare)}</p>

                <p>Target rows: {len(target\_compare)}</p>

                <p>Rows in common: {comparison.intersect\_rows}</p>

                <p>Rows only in source: {len(comparison.df1\_unq\_rows) if hasattr(comparison, 'df1\_unq\_rows') else 0}</p>

                <p>Rows only in target: {len(comparison.df2\_unq\_rows) if hasattr(comparison, 'df2\_unq\_rows') else 0}</p>

                <p>Columns match: <span class="{'pass' if comparison.all\_columns\_match() else 'fail'}">{comparison.all\_columns\_match()}</span></p>

                <p>All rows match: <span class="{'pass' if comparison.matches() else 'fail'}">{comparison.matches()}</span></p>

            </div>

            <div class="section">

                <h2>Column Statistics</h2>

                {comparison.column\_stats.to\_html() if hasattr(comparison, 'column\_stats') and isinstance(comparison.column\_stats, pd.DataFrame) else '<p>No column statistics available</p>'}

            </div>

            <div class="section">

                <h2>Sample Mismatches</h2>

                {pd.concat([

                    comparison.df1\_unq\_rows.assign(Source='Source Only').head(5) if hasattr(comparison, 'df1\_unq\_rows') and not comparison.df1\_unq\_rows.empty else pd.DataFrame(),

                    comparison.df2\_unq\_rows.assign(Source='Target Only').head(5) if hasattr(comparison, 'df2\_unq\_rows') and not comparison.df2\_unq\_rows.empty else pd.DataFrame()

                ]).to\_html() if (hasattr(comparison, 'df1\_unq\_rows') and hasattr(comparison, 'df2\_unq\_rows')) else '<p>No mismatches found</p>'}

            </div>

        </body>

        </html>

        """

        html\_output.write(html\_report.encode('utf-8'))

        html\_output.seek(0)

        output.seek(0)

        return output, html\_output

    except Exception as e:

        logger.error(f"Error generating DataCompy report: {str(e)}")

        raise Exception(f"Failed to generate DataCompy report: {str(e)}")

def generate\_ydata\_profile(source\_df: pd.DataFrame, target\_df: pd.DataFrame,

                         mapping\_df: pd.DataFrame) -> Tuple[BytesIO, BytesIO, BytesIO]:

    """

    Generate Y-Data Profiling reports including individual profiles and comparison.

    Args:

        source\_df: Source DataFrame

        target\_df: Target DataFrame

        mapping\_df: DataFrame containing column mapping information

    Returns:

        Tuple of (source\_profile, target\_profile, comparison\_profile) as BytesIO objects

    """

    try:

        # Create mapping dictionary from mapping\_df

        column\_mapping = dict(zip(

            mapping\_df['Source Column'],

            mapping\_df['Target Column']

        ))

        # Filter out unmapped and excluded columns

        excluded\_columns = mapping\_df[mapping\_df['Exclude from Comparison']]['Source Column'].tolist()

        valid\_columns = {

            src: tgt for src, tgt in column\_mapping.items()

            if tgt and not pd.isna(tgt) and src not in excluded\_columns

        }

        # Prepare DataFrames for comparison

        source\_cols = list(valid\_columns.keys())

        target\_cols = [valid\_columns[src] for src in source\_cols]

        source\_compare = source\_df[source\_cols].copy()

        target\_compare = target\_df[target\_cols].copy()

        # Rename target columns to match source columns for comparison

        target\_compare.columns = source\_cols

        # Convert problematic data types to string

        for col in source\_compare.columns:

            if source\_compare[col].dtype.name not in ['int64', 'float64', 'bool', 'datetime64[ns]', 'object']:

                source\_compare[col] = source\_compare[col].astype(str)

            if target\_compare[col].dtype.name not in ['int64', 'float64', 'bool', 'datetime64[ns]', 'object']:

                target\_compare[col] = target\_compare[col].astype(str)

        # Handle null values

        source\_compare = source\_compare.fillna(pd.NA)

        target\_compare = target\_compare.fillna(pd.NA)

        try:

            # Generate source profile with complete configuration

            source\_profile = ProfileReport(

                source\_compare,

                title="Source Data Profile",

                minimal=False,

                explorative=True,

                show\_variable\_description=True,

                correlations={

                    "auto": True,

                    "pearson": True,

                    "spearman": True,

                    "kendall": True,

                    "phi\_k": True,

                    "cramers": True,

                },

                html={

                    'style': {'full\_width': True},

                    'minify\_html': False

                },

                interactions={

                    "continuous": True

                },

                samples=None,

                missing\_diagrams={

                    "bar": True,

                    "matrix": True,

                    "heatmap": True,

                    "dendrogram": True

                },

                duplicates={

                    "head": 10

                },

                plot={

                    "correlation": {

                        "cmap": "RdBu",

                        "bad": "#000000"

                    },

                    "missing": True,

                    "histogram": {

                        "bins": 50

                    },

                    "scatter": True,

                    "value\_counts": True

                },

                vars={

                    "num": {

                        "quantiles": [0.05, 0.25, 0.5, 0.75, 0.95],

                        "chi\_squared\_threshold": 0.999

                    }

                }

            )

            # Generate target profile with same configuration

            target\_profile = ProfileReport(

                target\_compare,

                title="Target Data Profile",

                minimal=False,

                explorative=True,

                show\_variable\_description=True,

                correlations={

                    "auto": True,

                    "pearson": True,

                    "spearman": True,

                    "kendall": True,

                    "phi\_k": True,

                    "cramers": True,

                },

                html={

                    'style': {'full\_width': True},

                    'minify\_html': False

                },

                interactions={

                    "continuous": True

                },

                samples=None,

                missing\_diagrams={

                    "bar": True,

                    "matrix": True,

                    "heatmap": True,

                    "dendrogram": True

                },

                duplicates={

                    "head": 10

                },

                plot={

                    "correlation": {

                        "cmap": "RdBu",

                        "bad": "#000000"

                    },

                    "missing": True,

                    "histogram": {

                        "bins": 50

                    },

                    "scatter": True,

                    "value\_counts": True

                },

                vars={

                    "num": {

                        "quantiles": [0.05, 0.25, 0.5, 0.75, 0.95],

                        "chi\_squared\_threshold": 0.999

                    }

                }

            )

            # Generate comparison profile

            comparison\_profile = source\_profile.compare(target\_profile)

            # Create BytesIO objects for reports

            source\_output = BytesIO()

            target\_output = BytesIO()

            comparison\_output = BytesIO()

            try:

                # Generate complete HTML reports with all visualizations

                source\_profile.to\_file(source\_output, silent=True)

                target\_profile.to\_file(target\_output, silent=True)

                comparison\_profile.to\_file(comparison\_output, silent=True)

                # Reset buffer positions

                source\_output.seek(0)

                target\_output.seek(0)

                comparison\_output.seek(0)

                return source\_output, target\_output, comparison\_output

            except Exception as e:

                logger.error(f"Error generating profile reports: {str(e)}")

                raise Exception(f"Failed to generate complete profile reports: {str(e)}")

        except Exception as e:

            logger.error(f"Error in profile generation: {str(e)}")

            # Fallback to basic HTML reports

            source\_output = BytesIO()

            target\_output = BytesIO()

            comparison\_output = BytesIO()

            source\_report = f"""

            <html><head><title>Source Data Profile</title></head>

            <body><h1>Source Data Profile</h1>{source\_compare.describe().to\_html()}</body></html>

            """

            target\_report = f"""

            <html><head><title>Target Data Profile</title></head>

            <body><h1>Target Data Profile</h1>{target\_compare.describe().to\_html()}</body></html>

            """

            comparison\_report = f"""

            <html>

            <head><title>Data Comparison Report</title></head>

            <body>

            <h1>Data Comparison Report</h1>

            <h2>Source Data Summary</h2>{source\_compare.describe().to\_html()}

            <h2>Target Data Summary</h2>{target\_compare.describe().to\_html()}

            </body></html>

            """

            source\_output.write(source\_report.encode('utf-8'))

            target\_output.write(target\_report.encode('utf-8'))

            comparison\_output.write(comparison\_report.encode('utf-8'))

            source\_output.seek(0)

            target\_output.seek(0)

            comparison\_output.seek(0)

            return source\_output, target\_output, comparison\_output

    except Exception as e:

        logger.error(f"Error generating Y-Data profile: {str(e)}")

        raise Exception(f"Failed to generate Y-Data profile: {str(e)}")

def generate\_regression\_report(source\_df: pd.DataFrame, target\_df: pd.DataFrame,

                            mapping\_df: pd.DataFrame, dtype\_mapping: Dict[str, str]) -> BytesIO:

    """

    Generate Excel-based regression report with multiple tabs.

    Args:

        source\_df: Source DataFrame

        target\_df: Target DataFrame

        mapping\_df: DataFrame containing column mapping information

        dtype\_mapping: Dictionary mapping columns to their desired data types

    Returns:

        BytesIO object containing the report

    """

    try:

        output = BytesIO()

        with pd.ExcelWriter(output, engine='xlsxwriter') as writer:

            workbook = writer.book

            # Create formats for PASS/FAIL cells

            pass\_format = workbook.add\_format({'bg\_color': '#90EE90'})  # Light green

            fail\_format = workbook.add\_format({'bg\_color': '#FFB6C6'})  # Light pink

            # Generate AggregationCheck tab

            \_generate\_aggregation\_check(source\_df, target\_df, mapping\_df, writer,

                                     pass\_format, fail\_format)

            # Generate CountCheck tab

            \_generate\_count\_check(source\_df, target\_df, writer, pass\_format, fail\_format)

            # Generate DistinctCheck tab

            \_generate\_distinct\_check(source\_df, target\_df, mapping\_df, writer,

                                  pass\_format, fail\_format)

        output.seek(0)

        return output

    except Exception as e:

        logger.error(f"Error generating regression report: {str(e)}")

        raise Exception(f"Failed to generate regression report: {str(e)}")

def \_generate\_aggregation\_check(source\_df: pd.DataFrame, target\_df: pd.DataFrame,

                              mapping\_df: pd.DataFrame, writer: pd.ExcelWriter,

                              pass\_format: xlsxwriter.format.Format,

                              fail\_format: xlsxwriter.format.Format) -> None:

    """Generate the AggregationCheck tab in the regression report."""

    # Get numeric columns

    numeric\_cols = source\_df.select\_dtypes(include=[np.number]).columns

    results = []

    for col in numeric\_cols:

        if col in mapping\_df['Source Column'].values:

            target\_col = mapping\_df[mapping\_df['Source Column'] == col]['Target Column'].iloc[0]

            source\_sum = source\_df[col].sum()

            target\_sum = target\_df[target\_col].sum()

            match = np.isclose(source\_sum, target\_sum, rtol=1e-05)

            results.append({

                'Source Column': col,

                'Target Column': target\_col,

                'Source Sum': source\_sum,

                'Target Sum': target\_sum,

                'Result': 'PASS' if match else 'FAIL'

            })

    # Create DataFrame and write to Excel

    agg\_df = pd.DataFrame(results)

    agg\_df.to\_excel(writer, sheet\_name='AggregationCheck', index=False)

    # Apply conditional formatting

    worksheet = writer.sheets['AggregationCheck']

    result\_col = agg\_df.columns.get\_loc('Result')

    for row in range(len(agg\_df)):

        if agg\_df.iloc[row]['Result'] == 'PASS':

            worksheet.write(row + 1, result\_col, 'PASS', pass\_format)

        else:

            worksheet.write(row + 1, result\_col, 'FAIL', fail\_format)

def \_generate\_count\_check(source\_df: pd.DataFrame, target\_df: pd.DataFrame,

                         writer: pd.ExcelWriter,

                         pass\_format: xlsxwriter.format.Format,

                         fail\_format: xlsxwriter.format.Format) -> None:

    """Generate the CountCheck tab in the regression report."""

    count\_data = {

        'Source File Name': source\_df.name if hasattr(source\_df, 'name') else 'Source',

        'Target File Name': target\_df.name if hasattr(target\_df, 'name') else 'Target',

        'Source Count': len(source\_df),

        'Target Count': len(target\_df),

        'Result': 'PASS' if len(source\_df) == len(target\_df) else 'FAIL'

    }

    count\_df = pd.DataFrame([count\_data])

    count\_df.to\_excel(writer, sheet\_name='CountCheck', index=False)

    # Apply conditional formatting

    worksheet = writer.sheets['CountCheck']

    result\_col = count\_df.columns.get\_loc('Result')

    if count\_data['Result'] == 'PASS':

        worksheet.write(1, result\_col, 'PASS', pass\_format)

    else:

        worksheet.write(1, result\_col, 'FAIL', fail\_format)

def \_generate\_distinct\_check(source\_df: pd.DataFrame, target\_df: pd.DataFrame,

                           mapping\_df: pd.DataFrame, writer: pd.ExcelWriter,

                           pass\_format: xlsxwriter.format.Format,

                           fail\_format: xlsxwriter.format.Format) -> None:

    """Generate the DistinctCheck tab in the regression report."""

    # Get non-numeric columns

    non\_numeric\_cols = source\_df.select\_dtypes(exclude=[np.number]).columns

    results = []

    for col in non\_numeric\_cols:

        if col in mapping\_df['Source Column'].values:

            target\_col = mapping\_df[mapping\_df['Source Column'] == col]['Target Column'].iloc[0]

            source\_distinct = set(source\_df[col].dropna().unique())

            target\_distinct = set(target\_df[target\_col].dropna().unique())

            source\_count = len(source\_distinct)

            target\_count = len(target\_distinct)

            count\_match = source\_count == target\_count

            values\_match = source\_distinct == target\_distinct

            results.append({

                'Source Column': col,

                'Target Column': target\_col,

                'Source Distinct Count': source\_count,

                'Target Distinct Count': target\_count,

                'Count Match': 'PASS' if count\_match else 'FAIL',

                'Values Match': 'PASS' if values\_match else 'FAIL',

                'Source Distinct Values': ', '.join(map(str, sorted(source\_distinct))),

                'Target Distinct Values': ', '.join(map(str, sorted(target\_distinct)))

            })

    # Create DataFrame and write to Excel

    distinct\_df = pd.DataFrame(results)

    distinct\_df.to\_excel(writer, sheet\_name='DistinctCheck', index=False)

    # Apply conditional formatting

    worksheet = writer.sheets['DistinctCheck']

    count\_match\_col = distinct\_df.columns.get\_loc('Count Match')

    values\_match\_col = distinct\_df.columns.get\_loc('Values Match')

    for row in range(len(distinct\_df)):

        if distinct\_df.iloc[row]['Count Match'] == 'PASS':

            worksheet.write(row + 1, count\_match\_col, 'PASS', pass\_format)

        else:

            worksheet.write(row + 1, count\_match\_col, 'FAIL', fail\_format)

        if distinct\_df.iloc[row]['Values Match'] == 'PASS':

            worksheet.write(row + 1, values\_match\_col, 'PASS', pass\_format)

        else:

            worksheet.write(row + 1, values\_match\_col, 'FAIL', fail\_format)

def generate\_difference\_report(source\_df: pd.DataFrame, target\_df: pd.DataFrame,

                             join\_columns: List[str], mapping\_df: pd.DataFrame,

                             join\_mappings: Dict[str, str]) -> BytesIO:

    """

    Generate enhanced side-by-side difference report with highlighted differences.

    Args:

        source\_df: Source DataFrame

        target\_df: Target DataFrame

        join\_columns: List of source columns to join on

        mapping\_df: DataFrame containing column mapping information

        join\_mappings: Dictionary mapping source join columns to target join columns

    Returns:

        BytesIO object containing the report

    """

    try:

        output = BytesIO()

        # Create mapping dictionary from mapping\_df

        column\_mapping = dict(zip(

            mapping\_df['Source Column'],

            mapping\_df['Target Column']

        ))

        # Filter out unmapped and excluded columns

        excluded\_columns = mapping\_df[mapping\_df['Exclude from Comparison']]['Source Column'].tolist()

        valid\_columns = {

            src: tgt for src, tgt in column\_mapping.items()

            if tgt and not pd.isna(tgt) and src not in excluded\_columns

        }

        # Prepare DataFrames for comparison

        source\_cols = list(valid\_columns.keys())

        target\_cols = [valid\_columns[src] for src in source\_cols]

        source\_compare = source\_df[source\_cols].copy()

        target\_compare = target\_df[target\_cols].copy()

        # Get the target join column names

        target\_join\_columns = [join\_mappings[src] for src in join\_columns]

        # Merge datasets

        merged = pd.merge(

            source\_compare, target\_compare,

            left\_on=join\_columns,

            right\_on=target\_join\_columns,

            how='outer',

            suffixes=('\_source', '\_target'),

            indicator=True

        )

        with pd.ExcelWriter(output, engine='xlsxwriter') as writer:

            workbook = writer.book

            # Create formats

            diff\_format = workbook.add\_format({'bg\_color': '#FFB6C6'})  # Light pink

            header\_format = workbook.add\_format({

                'bold': True,

                'bg\_color': '#D3D3D3',

                'border': 1

            })

            # Process and write each category

            for category in ['left\_only', 'right\_only', 'both']:

                if category == 'left\_only':

                    sheet\_name = 'Source Only'

                    data = merged[merged['\_merge'] == 'left\_only']

                elif category == 'right\_only':

                    sheet\_name = 'Target Only'

                    data = merged[merged['\_merge'] == 'right\_only']

                else:

                    sheet\_name = 'Matching Records'

                    data = merged[merged['\_merge'] == 'both']

                if not data.empty:

                    # Write data to sheet

                    data.to\_excel(writer, sheet\_name=sheet\_name, index=False)

                    worksheet = writer.sheets[sheet\_name]

                    # Apply header format

                    for col\_num, value in enumerate(data.columns.values):

                        worksheet.write(0, col\_num, value, header\_format)

                    # For matching records, highlight differences

                    if category == 'both':

                        for row\_idx in range(len(data)):

                            for src\_col in source\_cols:

                                if src\_col not in join\_columns:  # Skip join columns

                                    tgt\_col = valid\_columns[src\_col]

                                    src\_val = data.iloc[row\_idx][f"{src\_col}\_source"]

                                    tgt\_val = data.iloc[row\_idx][f"{src\_col}\_target"]

                                    # Compare values and highlight differences

                                    if pd.notna(src\_val) and pd.notna(tgt\_val):

                                        if str(src\_val) != str(tgt\_val):

                                            col\_idx\_src = data.columns.get\_loc(f"{src\_col}\_source")

                                            col\_idx\_tgt = data.columns.get\_loc(f"{src\_col}\_target")

                                            worksheet.write(row\_idx + 1, col\_idx\_src, src\_val, diff\_format)

                                            worksheet.write(row\_idx + 1, col\_idx\_tgt, tgt\_val, diff\_format)

                    # Adjust column widths

                    for idx, col in enumerate(data.columns):

                        max\_length = max(

                            data[col].astype(str).apply(len).max(),

                            len(str(col))

                        )

                        worksheet.set\_column(idx, idx, max\_length + 2)

        output.seek(0)

        return output

    except Exception as e:

        logger.error(f"Error generating difference report: {str(e)}")

        raise Exception(f"Failed to generate difference report: {str(e)}")

def create\_individual\_reports\_zip(datacompy\_excel: BytesIO,

                                datacompy\_html: BytesIO,

                                ydata\_source\_report: BytesIO,

                                ydata\_target\_report: BytesIO,

                                ydata\_comparison\_report: BytesIO,

                                regression\_report: BytesIO,

                                difference\_report: BytesIO) -> BytesIO:

    """

    Create a ZIP file containing individual reports in separate folders.

    Args:

        datacompy\_report: DataCompy report as BytesIO

        ydata\_source\_report: Source data profile as BytesIO

        ydata\_target\_report: Target data profile as BytesIO

        ydata\_comparison\_report: Profile comparison report as BytesIO

        regression\_report: Regression report as BytesIO

        difference\_report: Difference report as BytesIO

    Returns:

        BytesIO object containing the ZIP file with individual reports

    """

    try:

        timestamp = datetime.now().strftime('%Y%m%d\_%H%M%S')

        output = BytesIO()

        with zipfile.ZipFile(output, 'w', zipfile.ZIP\_DEFLATED) as zf:

            # DataCompy Reports

            zf.writestr(f'datacompy/datacompy\_report\_{timestamp}.xlsx',

                       datacompy\_excel.getvalue())

            zf.writestr(f'datacompy/datacompy\_report\_{timestamp}.html',

                       datacompy\_html.getvalue())

            # Y-Data Profiles

            zf.writestr(f'ydata\_profile/source\_profile\_{timestamp}.html',

                       ydata\_source\_report.getvalue())

            zf.writestr(f'ydata\_profile/target\_profile\_{timestamp}.html',

                       ydata\_target\_report.getvalue())

            zf.writestr(f'ydata\_profile/comparison\_profile\_{timestamp}.html',

                       ydata\_comparison\_report.getvalue())

            # Regression Report

            zf.writestr(f'regression/regression\_report\_{timestamp}.xlsx',

                       regression\_report.getvalue())

            # Difference Report

            zf.writestr(f'differences/difference\_report\_{timestamp}.xlsx',

                       difference\_report.getvalue())

            # Add a README file

            readme\_content = """

Data Comparison Reports

1. datacompy/ - Contains detailed comparison report

2. ydata\_profile/ - Contains comprehensive data profiling

3. regression/ - Contains aggregation, count, and distinct value checks

4. differences/ - Contains side-by-side difference report

Generated on: {timestamp}

            """.format(timestamp=timestamp)

            zf.writestr('README.txt', readme\_content.strip())

        output.seek(0)

        return output

    except Exception as e:

        logger.error(f"Error creating individual reports zip: {str(e)}")

        raise Exception(f"Failed to create individual reports zip: {str(e)}")

def create\_consolidated\_report(datacompy\_excel: BytesIO,

                             datacompy\_html: BytesIO,

                             ydata\_source\_report: BytesIO,

                             ydata\_target\_report: BytesIO,

                             ydata\_comparison\_report: BytesIO,

                             regression\_report: BytesIO,

                             difference\_report: BytesIO) -> BytesIO:

    """

    Combine all reports into a single consolidated report.

    Args:

        datacompy\_report: DataCompy report as BytesIO

        ydata\_source\_report: Source data profile as BytesIO

        ydata\_target\_report: Target data profile as BytesIO

        ydata\_comparison\_report: Profile comparison report as BytesIO

        regression\_report: Regression report as BytesIO

        difference\_report: Difference report as BytesIO

    Returns:

        BytesIO object containing the consolidated ZIP file

    """

    try:

        timestamp = datetime.now().strftime('%Y%m%d\_%H%M%S')

        output = BytesIO()

        with zipfile.ZipFile(output, 'w', zipfile.ZIP\_DEFLATED) as zf:

            # Add all reports to a single consolidated file

            # DataCompy Reports

            zf.writestr(f'reports/datacompy\_report\_{timestamp}.xlsx',

                       datacompy\_excel.getvalue())

            zf.writestr(f'reports/datacompy\_report\_{timestamp}.html',

                       datacompy\_html.getvalue())

            zf.writestr(f'reports/ydata\_source\_profile\_{timestamp}.html',

                       ydata\_source\_report.getvalue())

            zf.writestr(f'reports/ydata\_target\_profile\_{timestamp}.html',

                       ydata\_target\_report.getvalue())

            zf.writestr(f'reports/ydata\_comparison\_profile\_{timestamp}.html',

                       ydata\_comparison\_report.getvalue())

            zf.writestr(f'reports/regression\_report\_{timestamp}.xlsx',

                       regression\_report.getvalue())

            zf.writestr(f'reports/difference\_report\_{timestamp}.xlsx',

                       difference\_report.getvalue())

            # Add a summary README

            readme\_content = """

Consolidated Data Comparison Report

This ZIP file contains the following reports:

1. datacompy\_report - Detailed comparison of datasets

2. ydata\_profile - Comprehensive data profiling

3. regression\_report - Aggregation, count, and distinct value checks

4. difference\_report - Side-by-side differences

Generated on: {timestamp}

            """.format(timestamp=timestamp)

            zf.writestr('README.txt', readme\_content.strip())

        output.seek(0)

        return output

    except Exception as e:

        logger.error(f"Error creating consolidated report: {str(e)}")

        raise Exception(f"Failed to create consolidated report: {str(e)}")

data\_reader.py

import pandas as pd

import numpy as np

import pyodbc

import requests

import json

import zipfile

from io import BytesIO

from sqlalchemy import create\_engine

import logging

from .utils import check\_file\_size

logger = logging.getLogger(\_\_name\_\_)

def read\_csv\_dat(file, delimiter=',', chunksize=None):

    """

    Read CSV or DAT file into a pandas DataFrame.

    Args:

        file: File object or path

        delimiter: Column separator

        chunksize: Number of rows to read at a time for large files

    Returns:

        pandas DataFrame

    """

    try:

        # Try to detect the encoding

        try:

            import chardet

            raw\_data = file.read()

            file.seek(0)  # Reset file pointer

            detected = chardet.detect(raw\_data)

            encoding = detected['encoding']

        except:

            encoding = 'utf-8'

        # First try with default settings

        try:

            if check\_file\_size(file) > 3 \* 1024 \* 1024 \* 1024:  # 3GB

                logger.info("Large file detected, using chunked reading")

                chunks = []

                for chunk in pd.read\_csv(file, delimiter=delimiter, chunksize=500000,

                                       encoding=encoding, on\_bad\_lines='skip'):

                    chunks.append(chunk)

                return pd.concat(chunks, ignore\_index=True)

            else:

                return pd.read\_csv(file, delimiter=delimiter, encoding=encoding,

                                 on\_bad\_lines='skip')

        except Exception as e:

            logger.warning(f"First attempt to read CSV failed: {str(e)}")

            # If first attempt fails, try with more flexible settings

            file.seek(0)  # Reset file pointer

            try:

                return pd.read\_csv(file, delimiter=delimiter, encoding=encoding,

                                 on\_bad\_lines='skip',  # Skip bad lines

                                 quoting=3,  # QUOTE\_NONE

                                 engine='python'  # More flexible but slower engine

                                )

            except Exception as e2:

                logger.error(f"Both attempts to read CSV failed. First error: {str(e)}, Second error: {str(e2)}")

                raise Exception(f"Failed to read file. Please check the delimiter and file format. Error: {str(e2)}")

    except Exception as e:

        logger.error(f"Error reading CSV/DAT file: {str(e)}")

        raise Exception(f"Failed to read file: {str(e)}")

def read\_sql(server, database, username=None, password=None, query=None):

    """

    Read data from SQL Server using a query.

    Args:

        server: SQL Server hostname

        database: Database name

        username: SQL Server username (optional for Windows auth)

        password: SQL Server password (optional for Windows auth)

        query: SQL query to execute

    Returns:

        pandas DataFrame

    """

    try:

        if not query:

            raise ValueError("Query cannot be empty")

        # Try different SQL Server drivers

        drivers = [

            'ODBC Driver 18 for SQL Server',

            'ODBC Driver 17 for SQL Server',

            'ODBC Driver 13 for SQL Server',

            'SQL Server',

            'SQL Server Native Client 11.0',

            'SQL Server Native Client 10.0'

        ]

        connection = None

        last\_error = None

        for driver in drivers:

            try:

                # Build connection string based on authentication method

                if username and password:

                    # SQL Server authentication

                    connection\_string = (

                        f"DRIVER={{{driver}}};"

                        f"SERVER={server};"

                        f"DATABASE={database};"

                        f"UID={username};"

                        f"PWD={password};"

                        "Trusted\_Connection=no;"

                    )

                else:

                    # Windows authentication

                    connection\_string = (

                        f"DRIVER={{{driver}}};"

                        f"SERVER={server};"

                        f"DATABASE={database};"

                        "Trusted\_Connection=yes;"

                        "TrustServerCertificate=yes;"

                    )

                connection = pyodbc.connect(connection\_string, timeout=30)

                logger.info(f"Successfully connected using driver: {driver}")

                break

            except Exception as e:

                last\_error = e

                logger.warning(f"Failed to connect using driver {driver}: {str(e)}")

                continue

        if connection is None:

            raise Exception(f"Failed to connect to SQL Server with any available driver. Last error: {str(last\_error)}")

        # Execute query in chunks

        chunks = []

        cursor = connection.cursor()

        # Execute the query

        cursor.execute(query)

        # Get column names

        columns = [column[0] for column in cursor.description]

        # Fetch data in chunks

        while True:

            rows = cursor.fetchmany(500000)

            if not rows:

                break

            chunk\_df = pd.DataFrame.from\_records(rows, columns=columns)

            chunks.append(chunk\_df)

        cursor.close()

        connection.close()

        return pd.concat(chunks, ignore\_index=True) if chunks else pd.DataFrame()

    except Exception as e:

        logger.error(f"Error reading from SQL Server: {str(e)}")

        raise Exception(f"Failed to read from SQL Server: {str(e)}")

def read\_stored\_proc(server, database, username=None, password=None, proc\_name=None):

    """

    Execute a stored procedure and return results.

    Args:

        server: SQL Server hostname

        database: Database name

        username: SQL Server username (optional for Windows auth)

        password: SQL Server password (optional for Windows auth)

        proc\_name: Name of the stored procedure

    Returns:

        pandas DataFrame

    """

    try:

        if not proc\_name:

            raise ValueError("Stored procedure name cannot be empty")

        # Try different SQL Server drivers

        drivers = [

            'ODBC Driver 18 for SQL Server',

            'ODBC Driver 17 for SQL Server',

            'ODBC Driver 13 for SQL Server',

            'SQL Server',

            'SQL Server Native Client 11.0'

        ]

        connection = None

        last\_error = None

        for driver in drivers:

            try:

                # Build connection string based on authentication method

                if username and password:

                    # SQL Server authentication

                    connection\_string = (

                        f"DRIVER={{{driver}}};"

                        f"SERVER={server};"

                        f"DATABASE={database};"

                        f"UID={username};"

                        f"PWD={password};"

                        "Trusted\_Connection=no;"

                    )

                else:

                    # Windows authentication

                    connection\_string = (

                        f"DRIVER={{{driver}}};"

                        f"SERVER={server};"

                        f"DATABASE={database};"

                        "Trusted\_Connection=yes;"

                        "TrustServerCertificate=yes;"

                    )

                connection = pyodbc.connect(connection\_string, timeout=30)

                logger.info(f"Successfully connected using driver: {driver}")

                break

            except Exception as e:

                last\_error = e

                logger.warning(f"Failed to connect using driver {driver}: {str(e)}")

                continue

        if connection is None:

            raise Exception(f"Failed to connect to SQL Server with any available driver. Last error: {str(last\_error)}")

        try:

            cursor = connection.cursor()

            # Execute stored procedure

            cursor.execute(f"EXEC {proc\_name}")

            # Fetch results

            columns = [column[0] for column in cursor.description]

            results = []

            while True:

                rows = cursor.fetchmany(500000)  # Fetch in chunks

                if not rows:

                    break

                results.extend(rows)

            return pd.DataFrame.from\_records(results, columns=columns)

        finally:

            cursor.close()

            connection.close()

    except Exception as e:

        logger.error(f"Error executing stored procedure: {str(e)}")

        raise Exception(f"Failed to execute stored procedure: {str(e)}")

def read\_teradata(server, database, username, password, query):

    """

    Read data from Teradata using a query.

    Args:

        server: Teradata server hostname

        database: Database name

        username: Teradata username

        password: Teradata password

        query: SQL query to execute

    Returns:

        pandas DataFrame

    """

    try:

        connection\_string = f"teradatasql://{username}:{password}@{server}/{database}"

        engine = create\_engine(connection\_string)

        # Execute query in chunks for large datasets

        chunks = []

        for chunk in pd.read\_sql(query, engine, chunksize=500000):

            chunks.append(chunk)

        return pd.concat(chunks, ignore\_index=True) if chunks else pd.DataFrame()

    except Exception as e:

        logger.error(f"Error reading from Teradata: {str(e)}")

        raise Exception(f"Failed to read from Teradata: {str(e)}")

def read\_api(url, method="GET", headers=None, body=None):

    """

    Read data from an API endpoint.

    Args:

        url: API endpoint URL

        method: HTTP method (GET or POST)

        headers: Request headers as dictionary

        body: Request body as dictionary

    Returns:

        pandas DataFrame

    """

    try:

        headers = json.loads(headers) if headers else {}

        body = json.loads(body) if body else {}

        if method.upper() == "GET":

            response = requests.get(url, headers=headers, params=body)

        else:

            response = requests.post(url, headers=headers, json=body)

        response.raise\_for\_status()

        data = response.json()

        # Handle different JSON structures

        if isinstance(data, list):

            return pd.DataFrame(data)

        elif isinstance(data, dict):

            # Try to find the data array in the response

            for key, value in data.items():

                if isinstance(value, list):

                    return pd.DataFrame(value)

            return pd.DataFrame([data])

        else:

            raise ValueError("Unexpected API response format")

    except Exception as e:

        logger.error(f"Error reading from API: {str(e)}")

        raise Exception(f"Failed to read from API: {str(e)}")

def read\_excel(file):

    """

    Read Excel file (xls/xlsx) into a pandas DataFrame.

    Args:

        file: File object or path

    Returns:

        pandas DataFrame

    """

    try:

        # Try to read with default settings first

        try:

            return pd.read\_excel(file)

        except Exception as first\_error:

            logger.warning(f"First attempt to read Excel failed: {str(first\_error)}")

            # If first attempt fails, try with more flexible settings

            file.seek(0)  # Reset file pointer

            try:

                return pd.read\_excel(

                    file,

                    engine='openpyxl',  # Try alternative engine

                    on\_bad\_lines='skip'  # Skip problematic rows

                )

            except Exception as second\_error:

                logger.error(f"Both attempts to read Excel failed. First error: {str(first\_error)}, Second error: {str(second\_error)}")

                raise Exception(f"Failed to read Excel file. Please check the file format. Error: {str(second\_error)}")

    except Exception as e:

        logger.error(f"Error reading Excel file: {str(e)}")

        raise Exception(f"Failed to read Excel file: {str(e)}")

def read\_parquet(file):

    """

    Read a Parquet file into a pandas DataFrame.

    Args:

        file: File object or path

    Returns:

        pandas DataFrame

    """

    try:

        return pd.read\_parquet(file)

    except Exception as e:

        logger.error(f"Error reading Parquet file: {str(e)}")

        raise Exception(f"Failed to read Parquet file: {str(e)}")

def read\_zipped\_files(zip\_file, delimiter=','):

    """

    Read flat files from a ZIP archive.

    Args:

        zip\_file: ZIP file object or path

        delimiter: Column separator for flat files

    Returns:

        pandas DataFrame

    """

    try:

        with zipfile.ZipFile(zip\_file) as z:

            # Get all file names in the ZIP

            flat\_files = [f for f in z.namelist() if f.endswith(('.csv', '.dat', '.txt'))]

            if not flat\_files:

                raise ValueError("No supported files found in ZIP archive")

            # Read and combine all files

            dfs = []

            for file in flat\_files:

                with z.open(file) as f:

                    # Convert to BytesIO for pandas to read

                    buffer = BytesIO(f.read())

                    df = read\_csv\_dat(buffer, delimiter=delimiter)

                    dfs.append(df)

            return pd.concat(dfs, ignore\_index=True) if dfs else pd.DataFrame()

    except Exception as e:

        logger.error(f"Error reading from ZIP archive: {str(e)}")

        raise Exception(f"Failed to read from ZIP archive: {str(e)}")

app.py

import streamlit as st

import pandas as pd

import numpy as np

from datetime import datetime, date, timezone

import os

from typing import Dict, List, Optional, Union, Any

import logging

from backend.data\_reader import (

    read\_csv\_dat,

    read\_sql,

    read\_stored\_proc,

    read\_teradata,

    read\_api,

    read\_parquet,

    read\_zipped\_files

)

from backend.mapping\_utils import (

    auto\_map\_columns,

    validate\_join\_columns,

    validate\_mapping,

    apply\_column\_mapping,

    get\_excluded\_columns

)

from backend.report\_generator import (

    generate\_datacompy\_report,

    generate\_ydata\_profile,

    generate\_regression\_report,

    generate\_difference\_report,

    create\_consolidated\_report,

    create\_individual\_reports\_zip

)

from backend.utils import check\_file\_size, setup\_logger

# Set page config

st.set\_page\_config(

    page\_title="Data Comparison Framework",

    page\_icon="📊",

    layout="wide"

)

# Set up logging

logger = logging.getLogger(\_\_name\_\_)

if not logger.handlers:

    logger = setup\_logger()

# Define the type mapping dictionary

TYPE\_MAPPING = {

    # Integer types

    'int': 'int64',

    'int32': 'int64',

    'int64': 'int64',

    'integer': 'int64',

    'numeric': 'int64',

    'bigint': 'int64',

    'smallint': 'int64',

    'tinyint': 'int64',

    # Float types

    'Float': 'float64',

    'float': 'float64',

    'float32': 'float64',

    'float64': 'float64',

    'decimal': 'float64',

    'double': 'float64',

    'real': 'float64',

    # String types

    'varchar': 'string',

    'nvarchar': 'string',

    'char': 'string',

    'nchar': 'string',

    'text': 'string',

    'ntext': 'string',

    'string': 'string',

    'object': 'string',

    # Date/Time types

    'date': 'datetime64[ns]',

    'datetime': 'datetime64[ns]',

    'datetime64': 'datetime64[ns]',

    'timestamp': 'datetime64[ns]',

    # Boolean types

    'Boolean': 'bool',

    'bool': 'bool',

    'bit': 'bool',

    # Default type

    'unknown': 'string'

}

def main():

    # Add header with image

    st.image("https://images.pexels.com/photos/414612/pexels-photo-414612.jpeg",

             caption="Data Comparison Dashboard",

             use\_column\_width=True)

    st.title("Data Comparison Framework")

    # Create sidebar for source and target selection

    with st.sidebar:

        st.header("Configuration")

        # Source selection

        st.subheader("Source Configuration")

        source\_type = st.selectbox(

            "Select Source Type",

            ["CSV file", "DAT file", "SQL Server", "Stored Procs",

             "Teradata", "API", "Parquet file", "Zipped Flat files"],

            key="source\_type"

        )

        # Target selection

        st.subheader("Target Configuration")

        target\_type = st.selectbox(

            "Select Target Type",

            ["CSV file", "DAT file", "SQL Server", "Stored Procs",

             "Teradata", "API", "Parquet file", "Zipped Flat files"],

            key="target\_type"

        )

        # Delimiter selection for file types

        if source\_type in ["CSV file", "DAT file", "Zipped Flat files"]:

            source\_delimiter = st.text\_input("Source Delimiter", value=",", key="source\_delimiter")

        if target\_type in ["CSV file", "DAT file", "Zipped Flat files"]:

            target\_delimiter = st.text\_input("Target Delimiter", value=",", key="target\_delimiter")

    # Main panel

    col1, col2 = st.columns(2)

    # Source data input

    with col1:

        st.header("Source Data")

        source\_data = handle\_data\_input("source", source\_type)

    # Target data input

    with col2:

        st.header("Target Data")

        target\_data = handle\_data\_input("target", target\_type)

    # If both source and target data are loaded

    if source\_data is not None and target\_data is not None:

        # Auto-map columns

        column\_mapping = auto\_map\_columns(source\_data.columns, target\_data.columns)

        st.header("Column Mapping")

            # Create initial mapping DataFrame with editable source and target columns

            mapping\_df = pd.DataFrame({

                'Source Column': list(column\_mapping.keys()),

                'Target Column': list(column\_mapping.values()),

                'Exclude from Comparison': [False] \* len(column\_mapping)

            })

            st.write("Edit column mappings below. You can modify both source and target columns.")

            # Create a selection widget for each mapping row

            edited\_mappings = []

            for idx, row in mapping\_df.iterrows():

                col1, col2, col3, col4 = st.columns([2, 2, 2, 1])

                with col1:

                    # Dropdown for source column selection

                    selected\_source = st.selectbox(

                        f"Source column",

                        options=[''] + list(source\_data.columns),

                        index=list(source\_data.columns).index(row['Source Column']) + 1 if row['Source Column'] in source\_data.columns else 0,

                        key=f"source\_col\_{idx}",

                        help="Select the source column"

                    )

                with col2:

                    st.markdown("→")  # Arrow indicating mapping direction

                with col3:

                    # Dropdown for target column selection

                    selected\_target = st.selectbox(

                        f"Target column",

                        options=[''] + list(target\_data.columns),

                        index=list(target\_data.columns).index(row['Target Column']) + 1 if row['Target Column'] in target\_data.columns else 0,

                        key=f"target\_col\_{idx}",

                        help="Select the corresponding target column"

                    )

                with col4:

                    # Checkbox for exclusion

                    exclude = st.checkbox(

                        "Exclude",

                        value=row['Exclude from Comparison'],

                        key=f"exclude\_{idx}",

                        help="Check to exclude this column from comparison"

                    )

                if selected\_source and selected\_target:  # Only add valid mappings

                    edited\_mappings.append({

                        'Source Column': selected\_source,

                        'Target Column': selected\_target,

                        'Exclude from Comparison': exclude

                    })

        # Convert edited mappings to DataFrame

        edited\_mapping = pd.DataFrame(edited\_mappings)

        # Create valid mappings dictionary (excluding None/empty values and excluded columns)

        valid\_mappings = {

            row['Source Column']: row['Target Column']

            for \_, row in edited\_mapping.iterrows()

            if row['Target Column'] and not pd.isna(row['Target Column']) and not row['Exclude from Comparison']

        }

        # Show mapping summary

        if valid\_mappings:

            st.success(f"✅ {len(valid\_mappings)} columns mapped successfully")

        else:

            st.warning("⚠️ No valid column mappings. Please map at least one column.")

        # Join columns selection

        st.subheader("Join Columns")

        # Create list of available join columns from valid mappings

        available\_columns = [(f"{src} → {tgt}", src)

                           for src, tgt in valid\_mappings.items()

                           if not edited\_mapping[edited\_mapping['Source Column'] == src]['Exclude from Comparison'].iloc[0]]

        if available\_columns:

            selected\_join\_columns = st.multiselect(

                "Select Join Columns",

                options=[col[0] for col in available\_columns],

                help="Select one or more columns to join the datasets"

            )

        else:

            st.warning("No valid column mappings available for join selection. Please map at least one column.")

            selected\_join\_columns = []

        # Data type mapping

        st.subheader("Data Type Mapping")

        dtype\_mapping = {}

        for col in source\_data.columns:

            # Get current column type

            current\_type = str(source\_data[col].dtype)

            # Map numpy/pandas types to our type mapping

            if 'int' in current\_type:

                base\_type = 'int'

            elif 'float' in current\_type:

                base\_type = 'Float'

            elif 'datetime' in current\_type:

                base\_type = 'datetime'

            elif 'bool' in current\_type:

                base\_type = 'Boolean'

            elif 'object' in current\_type:

                # Try to infer type from data

                sample = source\_data[col].dropna().head(100)

                if len(sample) > 0:

                    if all(isinstance(x, (int, np.integer)) for x in sample):

                        base\_type = 'int'

                    elif all(isinstance(x, (float, np.floating)) for x in sample):

                        base\_type = 'Float'

                    elif all(isinstance(x, bool) for x in sample):

                        base\_type = 'Boolean'

                    elif all(isinstance(x, (date, datetime, pd.Timestamp)) for x in sample):

                        base\_type = 'datetime'

                    else:

                        base\_type = 'varchar'

                else:

                    base\_type = 'varchar'

            else:

                base\_type = 'varchar'

            # Get mapped type from TYPE\_MAPPING

            mapped\_type = TYPE\_MAPPING.get(base\_type, TYPE\_MAPPING['varchar'])

            # Create options list from TYPE\_MAPPING values

            type\_options = sorted(set(TYPE\_MAPPING.values()))

            # Create the selectbox with proper type selection

            dtype\_mapping[col] = st.selectbox(

                f"Type for {col}",

                options=type\_options,

                index=type\_options.index(mapped\_type),

                help=f"Current type: {current\_type}"

            )

        # Validate mappings and show comparison button

        if valid\_mappings:

            # Validate data type compatibility

            is\_valid, error\_message = validate\_mapping(

                source\_data, target\_data,

                valid\_mappings,

                dtype\_mapping

            )

            if not is\_valid:

                st.error(f"❌ Mapping validation failed: {error\_message}")

                st.info("💡 Please review your column mappings and data types.")

            # Compare button

            compare\_button = st.button(

                "Compare",

                type="primary",

                disabled=not (selected\_join\_columns and is\_valid),

                help="Generate comparison reports for the mapped columns"

            )

            if compare\_button:

                try:

                    with st.spinner("Generating comparison reports..."):

                        progress\_text = st.empty()

                        progress\_bar = st.progress(0)

                        # Get the actual join columns and their mappings

                        progress\_text.text("Preparing join columns...")

                        join\_cols = []

                        join\_mappings = {}

                        for col\_pair in selected\_join\_columns:

                            src\_col = col\_pair.split(" → ")[0]

                            tgt\_col = valid\_mappings[src\_col]

                            join\_cols.append(src\_col)

                            join\_mappings[src\_col] = tgt\_col

                        progress\_bar.progress(10)

                        # Apply data type conversions

                        progress\_text.text("Converting data types...")

                        for col, dtype in dtype\_mapping.items():

                            # Skip if column is not in valid mappings

                            if col not in valid\_mappings:

                                continue

                            target\_col = valid\_mappings[col]

                            try:

                                if dtype == 'int64':

                                    source\_data[col] = pd.to\_numeric(source\_data[col], errors='coerce').astype('Int64')

                                    target\_data[target\_col] = pd.to\_numeric(target\_data[target\_col], errors='coerce').astype('Int64')

                                elif dtype == 'float64':

                                    source\_data[col] = pd.to\_numeric(source\_data[col], errors='coerce')

                                    target\_data[target\_col] = pd.to\_numeric(target\_data[target\_col], errors='coerce')

                                elif dtype == 'datetime64[ns]':

                                    source\_data[col] = pd.to\_datetime(source\_data[col], errors='coerce')

                                    target\_data[target\_col] = pd.to\_datetime(target\_data[target\_col], errors='coerce')

                                elif dtype == 'bool':

                                    source\_data[col] = source\_data[col].astype('boolean')

                                    target\_data[target\_col] = target\_data[target\_col].astype('boolean')

                                else:  # string type

                                    source\_data[col] = source\_data[col].astype(str)

                                    target\_data[target\_col] = target\_data[target\_col].astype(str)

                            except Exception as e:

                                st.error(f"❌ Error converting {col} to {dtype}: {str(e)}")

                                logger.error(f"Error converting column {col} to {dtype}: {str(e)}")

                                raise Exception(f"Data type conversion failed for column {col}: {str(e)}")

                        progress\_bar.progress(20)

                        # Generate reports with progress updates

                        progress\_text.text("Generating DataCompy report...")

                        datacompy\_excel, datacompy\_html = generate\_datacompy\_report(

                            source\_data, target\_data, join\_cols, edited\_mapping, join\_mappings

                        )

                        progress\_bar.progress(40)

                        progress\_text.text("Generating Y-Data Profiles...")

                        try:

                            ydata\_source\_report, ydata\_target\_report, ydata\_comparison\_report = generate\_ydata\_profile(

                                source\_data, target\_data, edited\_mapping

                            )

                        except Exception as e:

                            st.warning("⚠️ Y-Data profile generation failed, falling back to basic comparison")

                            logger.warning(f"Y-Data profile generation failed: {str(e)}")

                            # Create basic HTML reports as fallback

                            source\_output = BytesIO()

                            target\_output = BytesIO()

                            comparison\_output = BytesIO()

                            source\_report = f"""

                            <html><head><title>Source Data Profile</title></head>

                            <body><h1>Source Data Profile</h1>{source\_data.describe().to\_html()}</body></html>

                            """

                            target\_report = f"""

                            <html><head><title>Target Data Profile</title></head>

                            <body><h1>Target Data Profile</h1>{target\_data.describe().to\_html()}</body></html>

                            """

                            comparison\_report = f"""

                            <html>

                            <head><title>Data Comparison Report</title></head>

                            <body>

                            <h1>Data Comparison Report</h1>

                            <h2>Source Data Summary</h2>{source\_data.describe().to\_html()}

                            <h2>Target Data Summary</h2>{target\_data.describe().to\_html()}

                            </body></html>

                            """

                            source\_output.write(source\_report.encode('utf-8'))

                            target\_output.write(target\_report.encode('utf-8'))

                            comparison\_output.write(comparison\_report.encode('utf-8'))

                            source\_output.seek(0)

                            target\_output.seek(0)

                            comparison\_output.seek(0)

                            ydata\_source\_report = source\_output

                            ydata\_target\_report = target\_output

                            ydata\_comparison\_report = comparison\_output

                        progress\_bar.progress(60)

                        progress\_text.text("Generating Regression report...")

                        regression\_report = generate\_regression\_report(

                            source\_data, target\_data, edited\_mapping, dtype\_mapping

                        )

                        progress\_bar.progress(80)

                        progress\_text.text("Generating Difference report...")

                        difference\_report = generate\_difference\_report(

                            source\_data, target\_data, join\_cols, edited\_mapping, join\_mappings

                        )

                        progress\_bar.progress(90)

                        progress\_text.text("Creating final reports...")

                        consolidated\_report = create\_consolidated\_report(

                            datacompy\_excel,

                            datacompy\_html,

                            ydata\_source\_report,

                            ydata\_target\_report,

                            ydata\_comparison\_report,

                            regression\_report,

                            difference\_report

                        )

                        progress\_bar.progress(100)

                        progress\_text.text("✅ Comparison completed!")

                        # Display reports in UI

                        st.success("✅ Comparison completed successfully!")

                        # Create tabs for different reports

                        report\_tabs = st.tabs(["Y-Data Profiles", "DataCompy Report", "Downloads"])

                        with report\_tabs[0]:

                            ydata\_subtabs = st.tabs(["Source Profile", "Target Profile", "Comparison"])

                            with ydata\_subtabs[0]:

                                st.markdown("""

                                    <style>

                                        iframe {

                                            width: 100%;

                                            min-width: 1200px !important;

                                            height: 2000px !important;

                                            border: none;

                                        }

                                        .stMarkdown {

                                            max-width: 100% !important;

                                        }

                                        .element-container {

                                            width: 100% !important;

                                            max-width: none !important;

                                        }

                                    </style>

                                """, unsafe\_allow\_html=True)

                                st.components.v1.html(

                                    ydata\_source\_report.getvalue().decode(),

                                    height=2000,

                                    scrolling=True,

                                    width=1200

                                )

                            with ydata\_subtabs[1]:

                                st.components.v1.html(

                                    ydata\_target\_report.getvalue().decode(),

                                    height=2000,

                                    scrolling=True,

                                    width=1200

                                )

                            with ydata\_subtabs[2]:

                                st.components.v1.html(

                                    ydata\_comparison\_report.getvalue().decode(),

                                    height=2000,

                                    scrolling=True,

                                    width=1200

                                )

                        with report\_tabs[1]:

                            datacompy\_subtabs = st.tabs(["Excel View", "HTML View"])

                            with datacompy\_subtabs[0]:

                                try:

                                    # Show Excel report in tables

                                    excel\_data = pd.read\_excel(datacompy\_excel, sheet\_name=None)

                                    for sheet\_name, df in excel\_data.items():

                                        st.write(f"### {sheet\_name}")

                                        st.dataframe(df, use\_container\_width=True)

                                        st.markdown("---")

                                except Exception as e:

                                    st.error(f"Error displaying Excel report: {str(e)}")

                            with datacompy\_subtabs[1]:

                                # Show HTML report - use the already generated report

                                st.components.v1.html(datacompy\_html.getvalue().decode(), height=800, scrolling=True)

                        with report\_tabs[2]:

                            # Provide download links in columns

                            col1, col2 = st.columns(2)

                            with col1:

                                st.download\_button(

                                label="📥 Download Consolidated Report",

                                data=create\_consolidated\_report(

                                    datacompy\_excel,

                                    datacompy\_html,

                                    ydata\_source\_report,

                                    ydata\_target\_report,

                                    ydata\_comparison\_report,

                                    regression\_report,

                                    difference\_report

                                ),

                                file\_name=f"consolidated\_report\_{datetime.utcnow().strftime('%Y%m%d\_%H%M%S')}.zip",

                                mime="application/zip",

                                help="Download all reports in a single ZIP file"

                            )

                            with col2:

                                st.download\_button(

                                label="📊 Download Individual Reports",

                                data=create\_individual\_reports\_zip(

                                    datacompy\_excel,

                                    datacompy\_html,

                                    ydata\_source\_report,

                                    ydata\_target\_report,

                                    ydata\_comparison\_report,

                                    regression\_report,

                                    difference\_report

                                ),

                                file\_name=f"individual\_reports\_{datetime.utcnow().strftime('%Y%m%d\_%H%M%S')}.zip",

                                mime="application/zip",

                                help="Download reports as separate files in a ZIP"

                            )

                except Exception as e:

                    # Clear progress indicators

                    if 'progress\_text' in locals():

                        progress\_text.empty()

                    if 'progress\_bar' in locals():

                        progress\_bar.empty()

                    # Show detailed error message

                    error\_msg = str(e)

                    if "Failed to read file" in error\_msg:

                        st.error("❌ Error reading file. Please check the file format and delimiter.")

                        st.info("💡 Try selecting a different delimiter or check if the file is properly formatted.")

                    elif "SQL Server" in error\_msg:

                        st.error("❌ Database connection error. Please check your connection details.")

                        st.info("💡 Verify your server address, credentials, and ensure the database is accessible.")

                    elif "data type conversion" in error\_msg:

                        st.error("❌ Data type conversion error. Please review your data type mappings.")

                        st.info("💡 Some columns may contain values incompatible with the selected data types.")

                    else:

                        st.error(f"❌ Error during comparison: {error\_msg}")

                    # Log the full error

                    logger.error(f"Comparison error: {error\_msg}", exc\_info=True)

def handle\_data\_input(prefix: str, data\_type: str) -> Optional[pd.DataFrame]:

    """Handle different types of data input based on the selected type."""

    try:

            if data\_type in ["CSV file", "DAT file", "Parquet file", "Excel file"]:

                file\_types = {

                    "CSV file": ["csv"],

                    "DAT file": ["dat"],

                    "Parquet file": ["parquet"],

                    "Excel file": ["xls", "xlsx"]

                }

                uploaded\_file = st.file\_uploader(

                    f"Upload {data\_type}",

                    type=file\_types[data\_type],

                    key=f"{prefix}\_file"

                )

                if uploaded\_file:

                    if data\_type == "Parquet file":

                        return read\_parquet(uploaded\_file)

                    elif data\_type == "Excel file":

                        return read\_excel(uploaded\_file)

                    else:

                        delimiter = st.session\_state.get(f"{prefix}\_delimiter", ",")

                        return read\_csv\_dat(uploaded\_file, delimiter)

        elif data\_type in ["SQL Server", "Teradata", "Stored Procs"]:

            with st.expander(f"{data\_type} Connection Details"):

                server = st.text\_input("Server", key=f"{prefix}\_server")

                database = st.text\_input("Database", key=f"{prefix}\_database")

                auth\_method = st.radio(

                    "Authentication Method",

                    ["Windows Authentication", "SQL Server Authentication"],

                    key=f"{prefix}\_auth\_method",

                    help="Choose Windows Authentication to use your Windows credentials"

                )

                username = None

                password = None

                if auth\_method == "SQL Server Authentication":

                    username = st.text\_input("Username", key=f"{prefix}\_username")

                    password = st.text\_input("Password", type="password", key=f"{prefix}\_password")

                if data\_type == "Stored Procs":

                    proc\_name = st.text\_input("Stored Procedure Name", key=f"{prefix}\_proc")

                    if server and database and proc\_name and (auth\_method == "Windows Authentication" or (username and password)):

                        return read\_stored\_proc(server, database, username, password, proc\_name)

                else:

                    query = st.text\_area("SQL Query", key=f"{prefix}\_query")

                    if server and database and query and (auth\_method == "Windows Authentication" or (username and password)):

                        if data\_type == "SQL Server":

                            return read\_sql(server, database, username, password, query)

                        else:

                            return read\_teradata(server, database, username, password, query)

        elif data\_type == "API":

            with st.expander("API Details"):

                api\_url = st.text\_input("API URL", key=f"{prefix}\_api\_url")

                method = st.selectbox("Method", ["GET", "POST"], key=f"{prefix}\_api\_method")

                headers = st.text\_area("Headers (JSON)", key=f"{prefix}\_api\_headers")

                body = st.text\_area("Body (JSON)", key=f"{prefix}\_api\_body")

                if api\_url:

                    return read\_api(api\_url, method, headers, body)

        elif data\_type == "Zipped Flat files":

            uploaded\_zip = st.file\_uploader(

                "Upload ZIP file",

                type=["zip"],

                key=f"{prefix}\_zip"

            )

            if uploaded\_zip:

                delimiter = st.session\_state.get(f"{prefix}\_delimiter", ",")

                return read\_zipped\_files(uploaded\_zip, delimiter)

    except Exception as e:

        st.error(f"Error loading {prefix} data: {str(e)}")

        logger.error(f"Data loading error for {prefix}: {str(e)}")

        return None

if \_\_name\_\_ == "\_\_main\_\_":

    main()