**PySpark Coding Standards and Guidelines**

**1. Code Structure and Formatting**

1. **Consistent Indentation**: Use 4 spaces for indentation, no tabs.
2. **Max Line Length**: Keep line length to 79 characters (PEP8 standard).
3. **Blank Lines**: Use blank lines to separate logical sections of code.
4. **Class and Function Docstrings**: Use docstrings for all classes and functions to describe their purpose, parameters, and return values.
5. **Modular Functions**: Break large code blocks into functions for readability and reuse.
6. **Avoid Long Functions**: Keep functions short, ideally less than 50 lines.
7. **Variable Naming Conventions**: Use snake\_case for variable and function names (e.g., input\_data, compute\_stats).
8. **Class Naming Conventions**: Use CamelCase for class names (e.g., DataProcessor).
9. **Import Order**: Follow PEP8 import guidelines: standard libraries first, followed by third-party libraries, and local imports last.
10. **Group Imports**: Import all required modules at the beginning of the file.
11. **Avoid Star Imports**: Never use wildcard imports like from module import \*; import only necessary functions or classes.

**2. Performance Optimizations**

1. **Avoiding .collect()**: Avoid using collect() to move data to the driver unless absolutely necessary; this can overwhelm the driver.
2. **Use .cache() or .persist() Properly**: Use caching or persistence when you need to reuse a DataFrame, but clear cache when no longer needed.
3. **Partition Tuning**: Tune the number of partitions for large datasets; use .repartition() or .coalesce() to optimize.
4. **Filter Early**: Apply .filter() early to minimize the amount of data being processed.
5. **Avoid Wide Transformations**: Minimize operations like join, groupBy, distinct, as they can cause shuffles and slow down performance.
6. **Broadcast Joins**: Use broadcast() for small datasets in joins to avoid expensive shuffles.
7. **Avoid Redundant Actions**: Don’t perform multiple actions on the same RDD/DataFrame without persistence, as each action will recompute the entire lineage.
8. **Use DataFrame API**: Prefer the DataFrame API over RDDs, as it’s optimized for performance and easier to work with.
9. **Use Vectorized UDFs**: If possible, use Pandas UDFs (vectorized UDFs) as they are faster than regular PySpark UDFs.
10. If file zise is more than 2 gb don’t perform show() or broadcast()

**3. Code Efficiency and Readability**

1. **Chaining Operations**: Chain DataFrame operations where applicable, but avoid overly complex one-liners.
2. **Avoid Hardcoding Values**: Store constants in variables or configuration files instead of hardcoding them in the code.
3. **Use Meaningful Variable Names**: Variable names should clearly indicate the purpose of the variable.
4. **Minimal Use of Comments**: Comments should clarify *why* something is done, not *what* is being done if the code is already readable.
5. **Use Python’s Built-in Features**: Use built-in Python libraries like itertools, collections, and math to make code concise and efficient.

**4. Error Handling and Logging**

1. **Exception Handling**: Use try-except blocks to catch exceptions. Log errors and handle failures gracefully.
2. **Custom Exception Classes**: Define and use custom exceptions when needed.
3. **Logging Best Practices**: Use Python’s logging library instead of print statements for logging. Use different logging levels (DEBUG, INFO, WARNING, ERROR, CRITICAL).
4. **Avoid Silent Failures**: Ensure exceptions are handled or logged properly; avoid empty except clauses.
5. **Descriptive Error Messages**: Write descriptive error messages in raise and log.error statements.

**5. Scalability**

1. **Avoid Skewed Data**: Ensure that data is well-partitioned to avoid skewed processing times across nodes.
2. **Efficient Data Types**: Use efficient data types like IntegerType, LongType, StringType to optimize memory usage.
3. **Column Pruning**: Select only the columns you need from DataFrames; avoid processing unnecessary columns.
4. **Avoid Complex Aggregations**: Where possible, break complex aggregations into smaller steps or pre-aggregate data.
5. **Check Partition Sizes**: Ensure the number of partitions matches the cluster size, and that each partition has a reasonable size (~100-200 MB).
6. **Dynamic Allocation**: Enable Spark’s dynamic resource allocation to scale resources based on the workload.

**6. Data Management and File Handling**

1. **File Formats**: Prefer optimized file formats like Parquet or ORC over CSV for faster I/O.
2. **Partitioning by Keys**: Partition large datasets by keys (e.g., date, region) for faster retrieval and better data locality.
3. **Avoid Small Files**: Avoid creating too many small files when writing out data (use coalesce or repartition to reduce the number of output files).
4. **Schema Enforcement**: Always enforce schemas when reading data to avoid problems with inconsistent types.
5. **DataFrame Schema Management**: Use .printSchema() to verify the schema of DataFrames and ensure the data types are correct.

**7. Code Modularity and Reusability**

1. **DRY Principle (Don’t Repeat Yourself)**: Reuse code by creating utility functions or classes instead of duplicating logic.
2. **Use Configuration Files**: Store environment-specific configurations in external files (e.g., JSON, YAML) instead of hardcoding them in the script.
3. **Reusable UDFs**: Define commonly used UDFs as reusable functions instead of defining them inline multiple times.
4. **Modular Pipelines**: Break down large ETL processes into modular pipelines that can be tested and reused.

**8. Unit Testing and Validation**

1. **Unit Tests for UDFs**: Write unit tests for UDFs separately using unittest or pytest.
2. **Test Edge Cases**: Ensure your unit tests cover edge cases (e.g., null values, empty datasets).
3. **Mock External Dependencies**: Mock file systems, databases, and external APIs in tests to avoid external dependencies.
4. **Dataframe Comparisons in Tests**: Use .collect() and Python’s assert methods to compare DataFrames in unit tests.
5. **Continuous Integration**: Set up CI pipelines to automatically run tests when code is pushed to version control.

**9. Code Reviews and Collaboration**

1. **Use Code Reviews**: Always have another developer review your code to ensure quality.
2. **Version Control Best Practices**: Commit code frequently, write meaningful commit messages, and use branches for feature development.
3. **Documentation**: Document complex logic, functions, and classes using detailed docstrings and comments.

**10. PySpark-Specific Best Practices**

1. **Use SQL for Complex Queries**: Use Spark SQL for complex operations like window functions, joins, and aggregations when it improves readability and performance.
2. **Avoid RDDs for Most Operations**: The DataFrame API is more optimized than RDDs, so prefer it in most cases.
3. **Broadcast Variables**: Use sc.broadcast() for variables that are reused across multiple transformations and are not large.
4. **Accumulators for Debugging**: Use accumulators to track counters during transformations for debugging purposes.

**11. Security and Compliance**

1. **Sensitive Data Masking**: Mask sensitive data before writing to logs or output.
2. **Environment Variables**: Store credentials and sensitive information in environment variables or secret management systems, not in code.
3. **Valid database table names** : sample.emplyee , sample.id. any other combination apart from this throw error