Writing Data in Pyspark

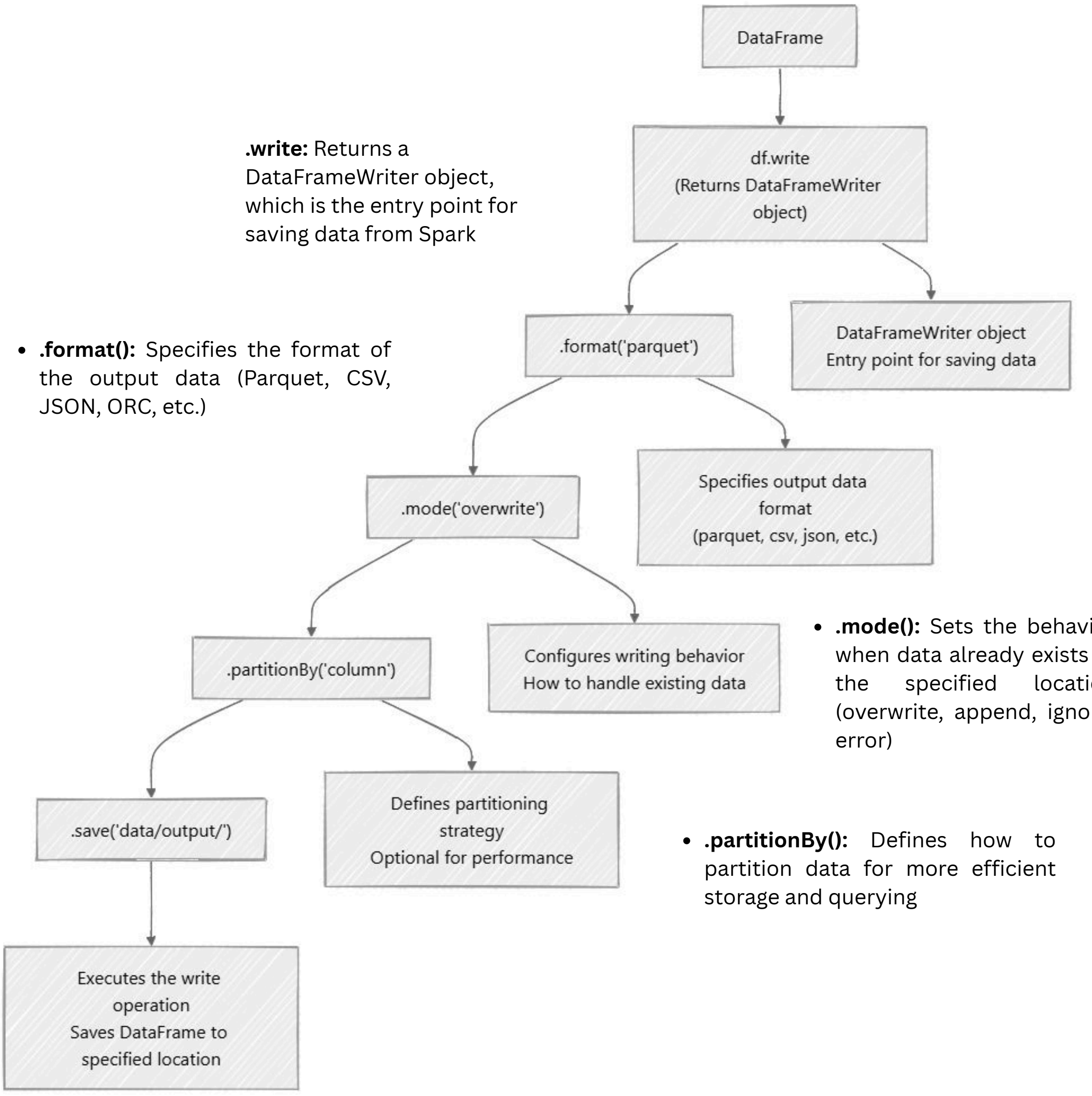
Writing Data

Explained Visually

https://manan.au

Writing Data in PySpark – Overview

.write: Returns a DataFrameWriter object, which is the entry point for saving data from Spark

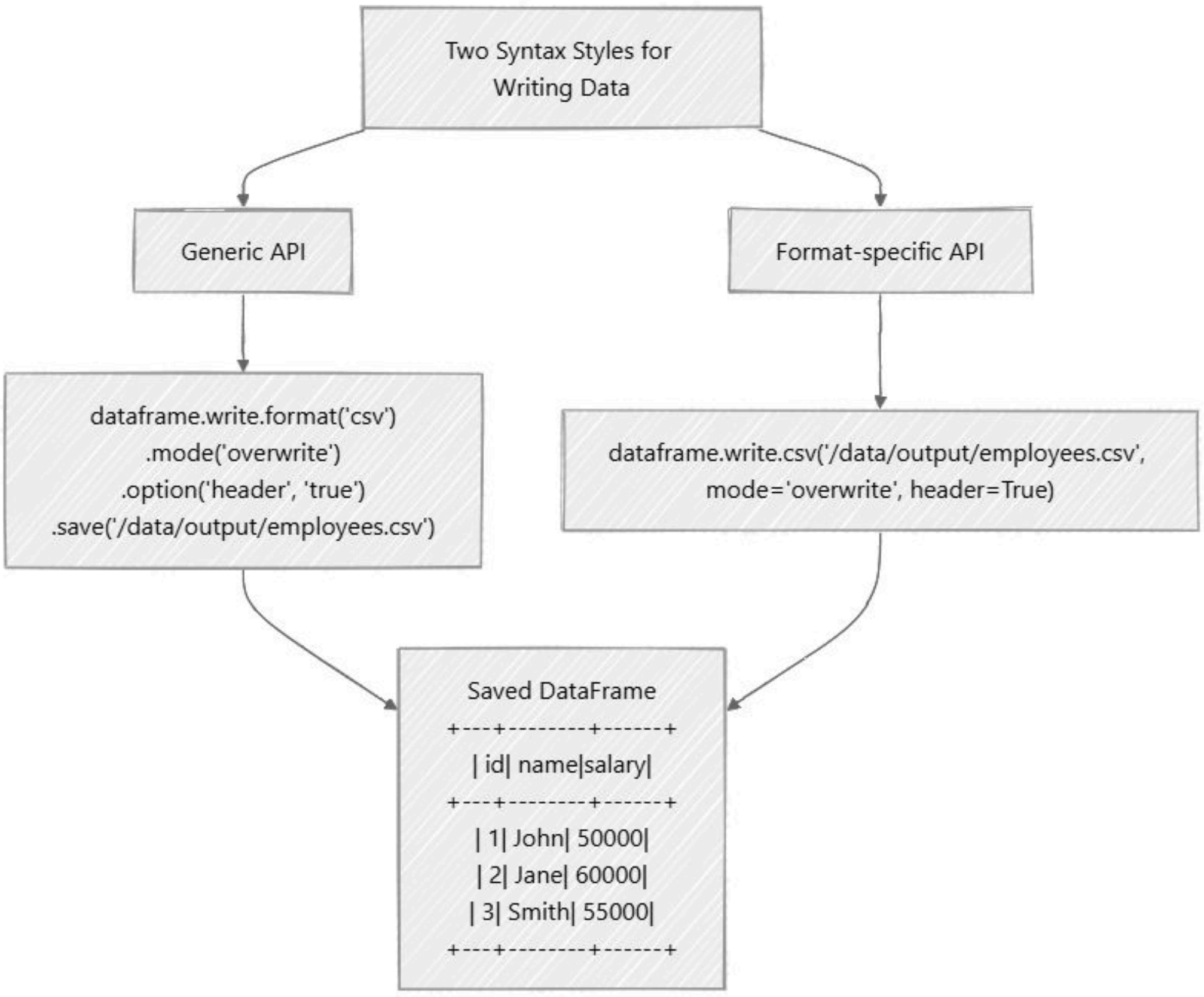


- **.format():** Specifies the format of the output data (Parquet, CSV, JSON, ORC, etc.)

- **.mode():** Sets the behavior when data already exists at the specified location (overwrite, append, ignore, error)

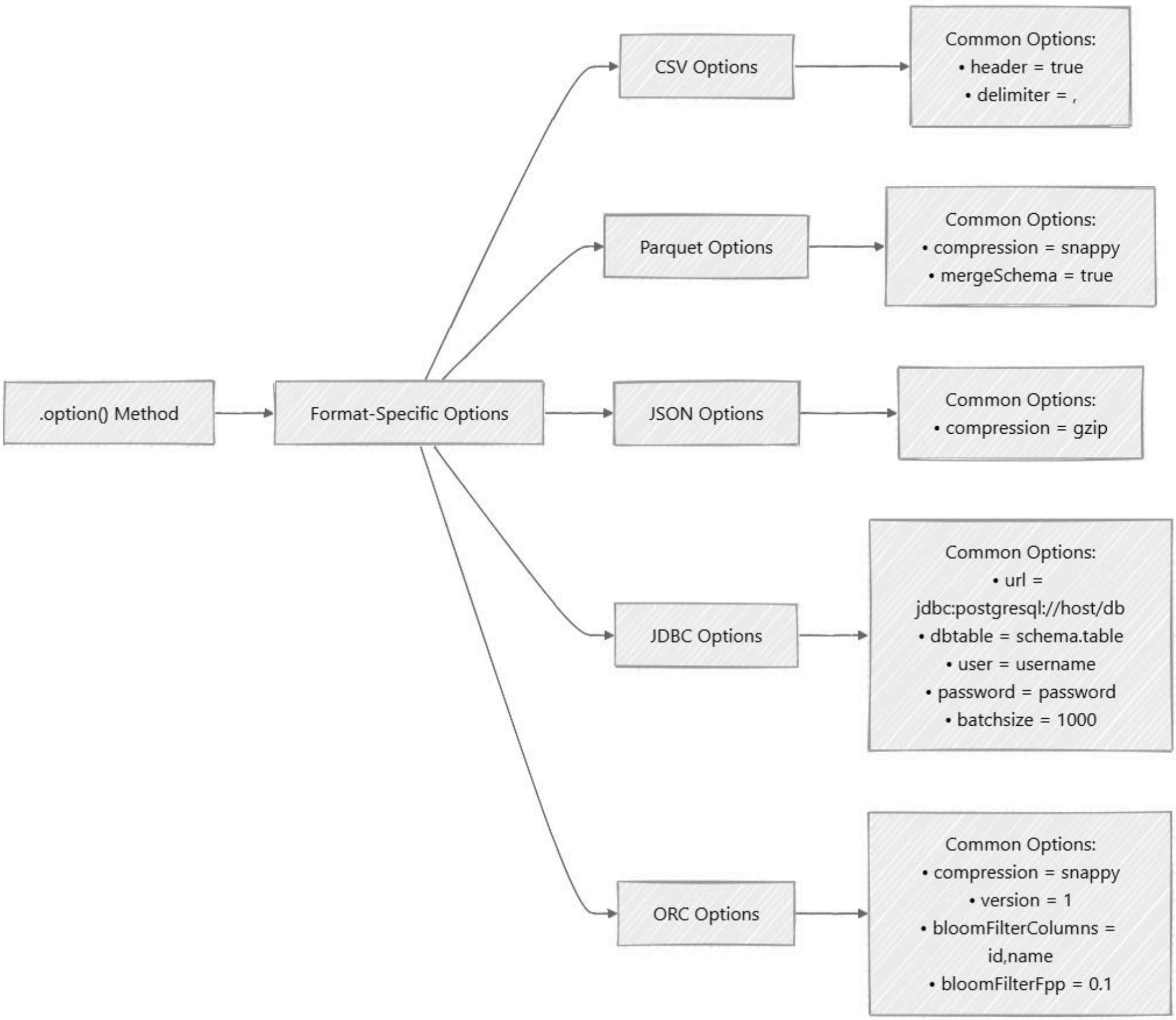
- **.partitionBy():** Defines how to partition data for more efficient storage and querying

Two Syntax Styles for Writing Data



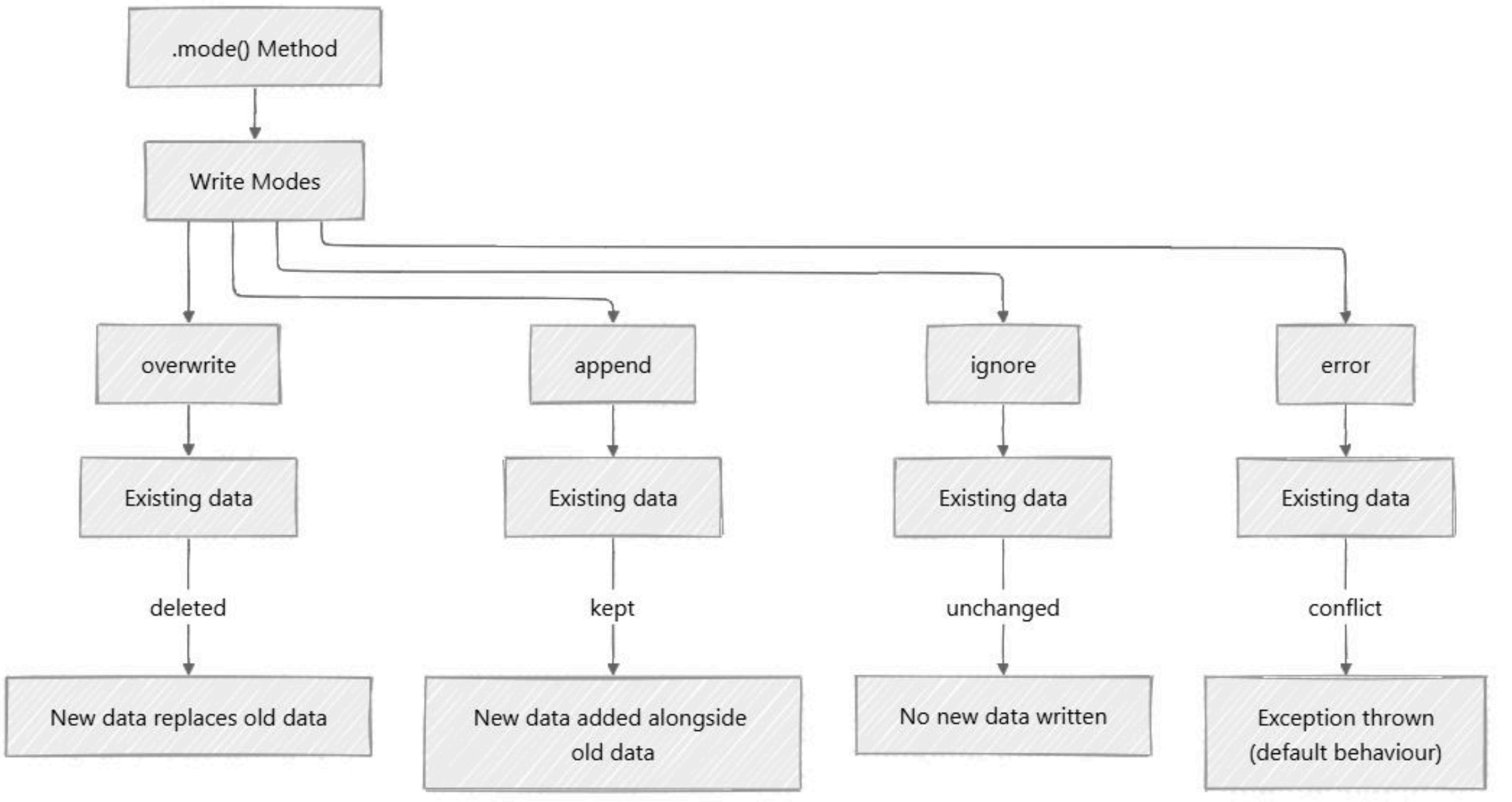
- Generic syntax uses format() to specify the file type explicitly
- Format-specific syntax provides shortcuts for common formats
- Format-specific methods improve code readability

The Option() Method for Data Writing



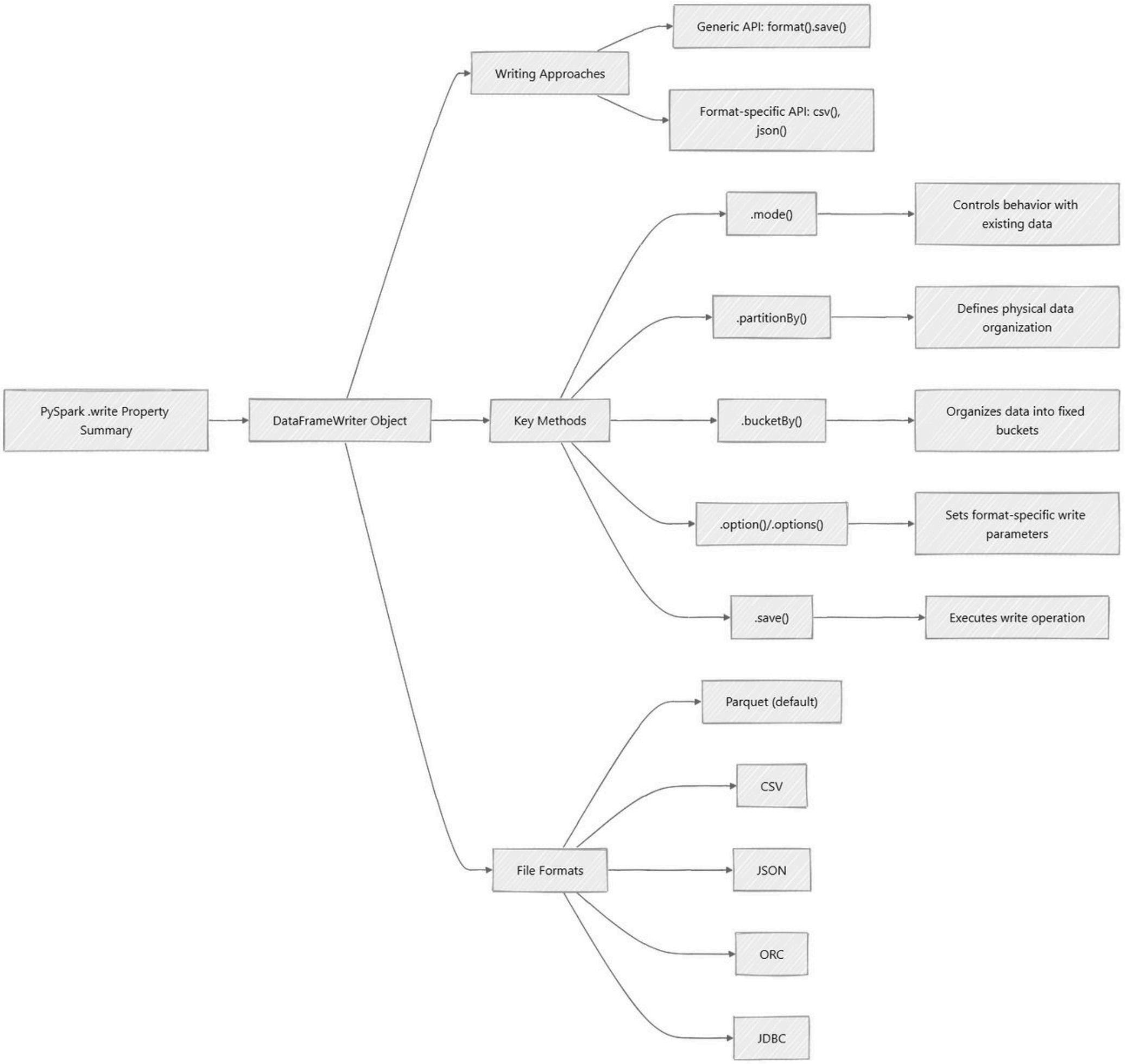
- Different file formats have their own specific options
- Options like 'header', 'delimiter', and 'quote' ensure proper data formatting
- Options like 'maxRecordsPerFile' help manage output file sizes
- Multiple options can be specified by chaining method calls

The .mode() method for specifying write Mode



- **Overwrite** mode completely replaces existing data with new data, useful for full refreshes
- **Append** mode adds new records alongside existing ones, ideal for incremental updates
- **Ignore** mode avoids writing when data already exists at the destination
- **Error** mode (default) fails the operation if data already exists, providing safe execution

Summary Writing Data



1. The .write property returns a DataFrameWriter object, which is the entry point for saving data from Spark.
2. Two syntax styles are available: generic (format().save()) and format-specific shortcuts (csv(), json(), etc.).
3. The .mode() method defines behavior when data already exists (overwrite, append, ignore, error).
4. Partitioning with .partitionBy() optimizes storage and query performance by organizing data based on column values.
5. Various file formats are supported with Parquet being the default format.