PySpark File Format Assignment

Part A — Load the Data

1. Load the CSV file using three methods:

- spark.read.csv("employee_data_200.csv")

- spark.read.format("csv").load("employee_data_200.csv")

- spark.read.option("header", "true").csv("employee_data_200.csv")

2. Load with header, inferSchema and delimiter options.

Part B — Convert the File

3. Convert CSV to JSON.

4. Convert CSV to Parquet.

5. Convert CSV to ORC.

6. Convert CSV to TEXT (first_name + last_name).

Part C — Transform and Save

7. Filter age > 30 → save Parquet.

8. Select columns → save JSON.

9. Add annual_salary column → save CSV.

10. Partition by department → save Parquet.

11. Repartition into 5 files → save.

12. Coalesce to 1 file → save.

Part D — Bad Records

13. Read CSV using:

- PERMISSIVE

- DROPMALFORMED

- FAILFAST

Part E — SQL Tasks

14. Create temp view employees and run:

- Count per department

- Avg salary per city

- Employees joined after 2020

Part F — Theory Questions

15. Differences: CSV, JSON, Parquet, ORC.

16. Why Parquet for big data?

17. repartition() vs coalesce()

18. What are bad records?