PySpark SQL Cheat Sheet

Setup & Register Table

```
df.createOrReplaceTempView("my_table")
```

Run SQL in PySpark

```
spark.sql("SELECT * FROM my_table WHERE age > 30").show()
```

Select and filter

Task	SQL Example
Select all columns	`SELECT * FROM my_table`
Select specific columns	`SELECT name, age FROM my_table`
Filter rows	`SELECT * FROM my_table WHERE age > 30`
Multiple conditions	`WHERE age > 30 AND country = 'US'`
IN clause	`WHERE country IN ('US', 'UK')`
LIKE pattern	`WHERE name LIKE 'A%'`
IS NULL	`WHERE email IS NULL`
IS NOT NULL	`WHERE email IS NOT NULL`

Aggregations

Task	SQL Example
Count total rows	`SELECT COUNT(*) FROM my_table`
Average value	`SELECT AVG(score) FROM my_table`
Group and count	`SELECT city, COUNT(*) FROM my_table GROUP BY city`
Group and aggregate multiple	`SELECT city, AVG(age), MAX(score) FROM my_table GROUP BY city`

Order and limit

Task	SQL Example
Sort ascending	`ORDER BY age ASC`
Sort descending	`ORDER BY age DESC`
Limit rows	`LIMIT 10`
Top-N per group *(windowing)*	*(requires PySpark Window, not SQL only)*

Date and time

Task	SQL Example
Extract year	`SELECT YEAR(date_column) FROM my_table`
Extract month	`SELECT MONTH(date_column)`
Date difference	`DATEDIFF(end_date, start_date)`
Current date	`CURRENT_DATE()`
Format date	`DATE_FORMAT(date_column, 'yyyy-MM')`

Joins

Task	SQL Example
Inner join	`SELECT * FROM a JOIN b ON a.id = b.id`
Left join	`SELECT * FROM a LEFT JOIN b ON a.id = b.id`
Right join	`SELECT * FROM a RIGHT JOIN b ON a.id = b.id`
Full outer join	`SELECT * FROM a FULL JOIN b ON a.id = b.id`

Case

Task	SQL Example
CASE WHEN	`SELECT name, CASE WHEN age > 18 THEN 'adult' ELSE 'minor' END AS status FROM my_table`
COALESCE	`SELECT COALESCE(email, 'no-email')`
NULLIF	`NULLIF(col1, col2)`