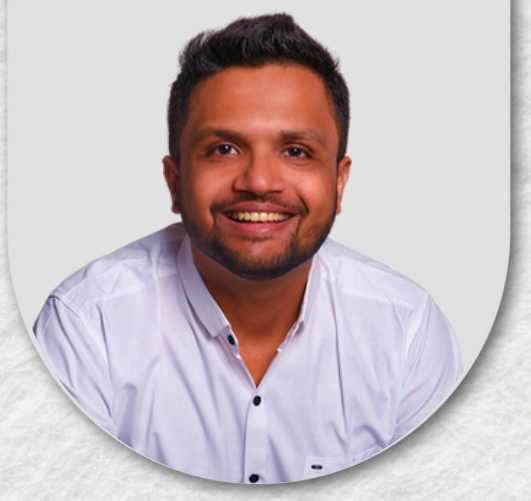


PySpark

Window Functions in PySpark – Cheat Sheet

Abhishek Agarwal
Data Engineer

Function	Description	Code Example
row_number()	Unique row number per window partition	row_number().over(Window.partitionBy("category"))
rank()	Ranking with gaps	rank().over(Window.partitionBy("region"))
dense_rank()	Ranking without gaps	dense_rank().over(Window.partitionBy("region"))
ntile(n)	Divides rows into n buckets	ntile(4).over(Window.orderBy("sales"))
lag()	Value from previous row	lag("revenue", 1).over(Window.partitionBy("id"))

Function	Description	Code Example
lead()	Value from next row	lead("revenue", 1).over(Window.partitionBy("id"))
first()	First value in partition	first("value").over(Window.partitionBy("group"))
last()	Last value in partition	last("value").over(Window.partitionBy("group"))
sum()	Running total	sum("amount").over(Window.orderBy("date"))
avg()	Running average	avg("score").over(Window.partitionBy("student"))

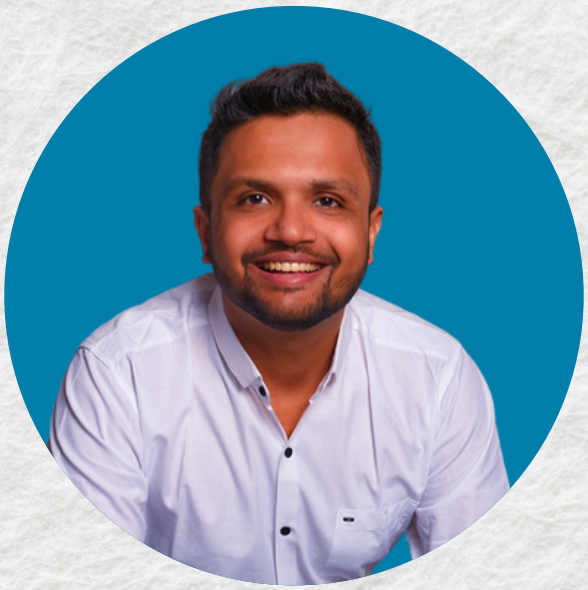
Function	Description	Code Example
min() / max()	Min/Max in window	max("score").over(Window.partitionBy("subject"))

Remember to import:

- from pyspark.sql.window import Window
- from pyspark.sql.functions import *

Use partitionBy() to group data and orderBy() to define row order within the window.

Follow for more content like this



Abhishek Agarwal

Azure Data Engineer