Detailed PySpark Topics – Transformations & Inbuilt Functions

Below is a well-organized list of **every important PySpark transformation + action + built-in function category**, with explanation and subtopics.

---

1. PySpark DataFrame Basics

Before transformations, teach:

- What is a DataFrame?
- Lazy evaluation
- Narrow vs Wide transformations
- Actions vs Transformations

---

2. Data Loading & Basic Operations

- Creating DataFrames (from CSV, JSON, Parquet)
- Print schema, describe, summary
- show(), head(), collect(), count()

---

3. Column Operations

Transformations

- select()
- selectExpr()
- withColumn()
- withColumnRenamed()
- drop(), dropDuplicates()

Column expressions

- lit()
- col()
- expr()

4. Filtering & Conditional Logic

Transformations

- filter() / where()
- between()
- isin()
- like(), rlike()

Conditional functions

- when(), otherwise()
- case-when using expr()

5. Handling NULL Values

Functions

- isNull(), isNotNull()
- fillna()
- dropna()
- na.replace()

6. String Functions

Most-used string operations:

- lower(), upper()
- trim(), ltrim(), rtrim()
- length()
- substring()
- split()
- concat(), concat_ws()
- regexp_replace(), regexp_extract()
- translate()

7. Date & Timestamp Functions

Teach very clearly:

- current_date(), current_timestamp()
- date_format()
- to_date(), to_timestamp()
- datediff(), months_between()
- add_months(), date_add(), date_sub()
- year(), month(), dayofmonth(), weekofyear()

---

8. Numeric Functions

- abs(), round(), floor(), ceil()
- pow(), sqrt()
- greatest(), least()

---

9. Array Functions

Important for complex JSON:

- array(), array_contains()
- explode()
- size()
- sort_array()
- array_distinct()
- arrays_zip()

---

10. Struct Functions

Teach nested fields:

- struct()
- getField()
- withField()
- renaming nested fields
- dot notation (col("a.b.c"))

11. Map Functions

- create_map()
- map_keys(), map_values()
- element_at()

12. Aggregation & Grouping

Transformations

- groupBy()
- rollup(), cube()

Functions

- sum(), avg(), min(), max(), count()
- countDistinct()
- collect_list(), collect_set()
- agg()

13. Joins

Teach all join types:

- inner
- left, right, full
- left semi, left anti
- cross join
- broadcast join (very important)

14. Window Functions (Must Teach)

Very important for analytics:

- row_number()
- rank(), dense_rank()

- lag(), lead()
- cumulative sum: sum().over(window)
- window specifications (partitionBy, orderBy, rowsBetween)

---

15. Repartitioning & Optimization

- repartition()
- coalesce()
- partitionBy() in write
- cache(), persist()
- checkpoint()

---

16. Reading/Writing Files

- read/write CSV, JSON, Parquet, Delta
- mode("append"), mode("overwrite")
- writing partitioned data
- saveAsTable()

---

17. Complex JSON Handling

- from_json()
- to_json()
- schema inference vs manual schema
- explode nested JSON arrays

---

18. UDFs

- What is a UDF?
- Normal UDF
- Pandas UDF (vectorized, faster)
- When **NOT** to use UDF (performance)

---

19. Actions (To Trigger Transformations)

- show()
- collect()
- count()
- take()
- foreach()

Bonus Topics (Optional for Freshers)

- Spark SQL using createOrReplaceTempView
