PySpark Interview Cheat Sheet

ABHINAV SINGH

1. Initialization

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("example").getOrCreate()
```

2. Loading Data

```
df = spark.read.format("csv").option("header", "true").load("path/to/file.csv")
```

3. Rename Column

```
df = df.withColumnRenamed("old_column", "new_column")
```

4. Rename Multiple Columns

```
df = df.withColumnRenamed("old_column1",
"new_column1").withColumnRenamed("old_column2", "new_column2")
```

5. Add Column

```
df = df.withColumn("new_column", lit("value"))
```

6. Drop Column

```
df = df.drop("column")
```

ABHINAV SINGH

7. Select Columns

```
df.select("column1", "column2")
```

8. Filter Rows

```
df.filter(df["column"] == "value")
```

9. Sort Rows

```
df.orderBy(df["column"].asc())
df.orderBy(df["column"].desc()
```

10. Remove Duplicates

```
df.dropDuplicates()
df.dropDuplicates(['column'])
```

11. Union

```
df1.union(df2)
```

12. Conditional Logic

```
df.withColumn("new_column", when(df["column"] == "value",
"result").otherwise("other_result"))
```

ABHINAV SINGH

13. Contains

```
df.filter(df["column"].contains("value"))
```

14. Summary Statistics

```
df.describe()
```

15. Trim

```
df.select(trim(df["column"]))
```

16. Joins

```
df1.join(df2, df1["key_column"] == df2["key_column"], "inner")
```

17. Aggregate Functions

```
df.groupBy("column").agg(count("*"), sum("column"))
```

18. Window Functions

```
from pyspark.sql.window import Window
windowSpec = Window.partitionBy("column").orderBy("column")
```

ABHINAV SINGH

19. Running Total

```
df.withColumn("running_total", sum("column").over(windowSpec))
```

20. Rank

```
df.withColumn("rank", rank().over(windowSpec))
```

21. Dense Rank

```
df.withColumn("dense_rank", dense_rank().over(windowSpec))
```

22. Repartition

```
df.repartition(6)
```

23. Coalesce

```
df.coalesce(6)
```

24. Partition

```
df.write.partitionBy("column").mode("overwrite").save("path")
```

ABHINAV SINGH

25. Bucketing

```
df.write.bucketBy(4, "column").sortBy("column").saveAsTable("table_name")
```

26. Cast Column

```
df.withColumn("column", df["column"].cast("new_type"))
```

27. Fill Nulls

```
df.fillna("value")
df.fillna({"column": "value"})
```

28. Literal

```
from pyspark.sql.functions import lit
df.select(lit(1).alias("LiteralCol"))
```

29. GroupBy

```
df.groupBy("column").count()
```

30. Pivot

```
df.groupBy("column").pivot("pivot_column").sum("value_column")
```

ABHINAV SINGH

31. Date Functions

```
df.select(current_date(), current_timestamp())
```

32. Replace Values

```
df.replace("old_value", "new_value")
```

33. Drop Rows with Nulls

```
df.na.drop()
```

34. Regex Functions

```
df.select(regexp_replace(col("col"), "pattern", "replacement"))
```

35. Drop Duplicates

```
df.groupBy("column").df.dropDuplicates(subset=["column1", "column2"])
```

ABHINAV SINGH