

# repartition and file layout

---

## Key rule

```
df.repartition(n, "fiscal_year")
```

- All rows with the same **fiscal\_year** go to **exactly one Spark partition**
- With **.write.partitionBy("fiscal\_year")**:
  - **≤1 file per fiscal\_year per write**

## Case Study (3 fiscal years, 20 original partitions)

- **No repartition** (20 partitions)
  - Each partition may contain mixed years
  - Each partition writes its own file
  - → 20 files where each could contain records from any year
- **.repartition(3, "fiscal\_year")**
  - 3 Spark partitions
  - Each year → one partition
  - → 3 files, 1 file per year
- **.repartition(2, "fiscal\_year")**
  - One partition holds 1 year
  - One partition holds 2 years
  - → 2 files, 1 file contains 2 years
- **.repartition(4, "fiscal\_year")**
  - 4 partitions, 3 keys
  - 1 partition empty
  - → 4 files, 1 file is empty