spark-shell

If a df is cached but need to re-read it from HDFS path because the underlying parquet files have been updated, unpersist it first. Otherwise the df won't be refreshed. E.g.,

```
var df = spark.read.parquet("xxx")
df.cache()
df.unpersist()
var df = spark.read.parquet("xxx")
```

Join

To discard duplicate join key columns:

```
df1.join(df2, Seq("user_id"), "left") // only df1.user_id will be kept
```

To keep join key columns in both dfs:

```
df1.join(df2, $"df1.user_id" === $"df2.user_id", "left") // both df1.user_id and df2.user_id will be kept, so can use df2.user_id later, e.g., in .select($"df2.user_id")
```

Partition

To find size of each partition of df (see here):

```
df.rdd.mapPartitions(iter => Iterator(iter.size)).collect().mkString(", ")
```

Repartition after filtering.