13 Initial number of partitions

The maximum file size by configuration will be 128 mb.

`spark.sql.files.maxPartitionBytes` (default ~128 MB)

spark.sparkContext.defaultParallelism = number of cores

Number of partitions will be → max(number of cores , filesize/128mb)

Initial number of partitions when we are reading multiple small files .

```
spark.conf.get("spark.sql.files.openCostInBytes") # (default ~4 MB)

from pyspark.sql.types import *
from pyspark.sql.functions import *
orders_schema = StructType(
[
StructField("order_id",IntegerType(),False)
,StructField("customer_id",IntegerType(),False)
,StructField("product_id",IntegerType(),False)
,StructField("unit_price",FloatType(),False)
,StructField("order_date",DateType(),False)
,StructField("order_status",StringType(),False)
,StructField("state",StringType(),False)
,StructField("quantity",IntegerType(),False)
]
)

df_orders = spark.read.csv("/data/orders_600mb.csv" , schema = orders_schema , header = True , inferSchema=False)
#df = orders_df.join(products_df,"product_id", "inner")
#orders_df.groupBy("order_status").agg(count("*")).write.format("noop").
```

```
mode("overwrite").save()
df_orders.repartition(50).write.mode("overwrite").format("csv").save("/data/multiple_files/")

df = spark.read.csv("/data/multiple_files")
df.rdd.getNumPartitions()
```