1

```
from pyspark.sql.types import StructField, StructType, StringType, IntegerType
```

2

```
my_schema = StructType(
    [
        StructField("DEST_COUNTRY_NAME", StringType(), True),
        StructField("DESTORIGIN_COUNTRY_NAME_COUNTRY_NAME", StringType(), True),
        StructField("count", IntegerType(), True)
    ]
)
```

3

```
file_location = "/FileStore/tables/2015_summary.csv"

df = spark.read.format("csv") \
  .option("inferSchema", "false") \
  .option("header", "false") \
  .option("skiprows", "1") \
  .schema(my_schema)\
  .option("mode", "PERMISSIVE")\
  .load(file_location)

df.show(2)
```

▸ df: pyspark.sql.dataframe.DataFrame = [DEST_COUNTRY_NAME: string, DESTORIGIN_COUNTRY_NAME_COUNTRY_NAME: string ... 1 more field]

```
+-----------------+-----------------------------------+-----+
|DEST_COUNTRY_NAME|DESTORIGIN_COUNTRY_NAME_COUNTRY_NAME|count|
+-----------------+-----------------------------------+-----+
|    United States|                            Romania|   15|
|    United States|                            Croatia|    1|
+-----------------+-----------------------------------+-----+
only showing top 2 rows
```

4

```
%fs
ls /FileStore/tables/
```

Table

	path	name	size	modificationTime
1	dbfs:/FileStore/tables/2015_summary-1.csv	2015_summary-1.csv	7080	1733968845000
2	dbfs:/FileStore/tables/2015_summary.csv	2015_summary.csv	7080	1733886028000

2 rows

corrupted record

6

```
%fs
ls /FileStore/tables/
```

Table

	path	name	size	modificationTime
1	dbfs:/FileStore/tables/2015_summary-1.csv	2015_summary-1.csv	7080	1733968845000
2	dbfs:/FileStore/tables/2015_summary.csv	2015_summary.csv	7080	1733886028000
3	dbfs:/FileStore/tables/employees.csv	employees.csv	225	1733971046000

3 rows

7

```
file_location = "/FileStore/tables/employees.csv"

df = spark.read.format("csv") \
  .option("inferSchema", "true") \
  .option("header", "true") \
  .option("mode", "PERMISSIVE")\
  .load(file_location)

df.show() ## ignore the nominee corrupted record 3-4
```

df: pyspark.sql.dataframe.DataFrame = [id: integer, name: string ... 4 more fields]

```
+---+--------+---+------+------------+--------+
| id|    name|age|salary|     address| nominee|
+---+--------+---+------+------------+--------+
|  1|  Manish| 26| 75000|       bihar|nominee1|
|  2|  Nikita| 23|100000|uttarpradesh|nominee2|
|  3|  Pritam| 22|150000|   Bangalore|   India|
|  4|Prantosh| 17|200000|     Kolkata|   India|
|  5|  Vikash| 31|300000|        null|nominee5|
+---+--------+---+------+------------+--------+
```

8

```
df = spark.read.format("csv") \
  .option("inferSchema", "true") \
  .option("header", "true") \
  .option("mode", "DROPMALFORMED")\
  .load(file_location)

df.show()
```

df: pyspark.sql.dataframe.DataFrame = [id: integer, name: string ... 4 more fields]

```
+---+------+---+------+------------+--------+
| id|  name|age|salary|     address| nominee|
+---+------+---+------+------------+--------+
|  1|Manish| 26| 75000|       bihar|nominee1|
|  2|Nikita| 23|100000|uttarpradesh|nominee2|
|  5|Vikash| 31|300000|        null|nominee5|
+---+------+---+------+------------+--------+
```

9

```
1  df = spark.read.format("csv") \
2    .option("inferSchema", "true") \
3    .option("header", "true") \
4    .option("mode", "FAILFAST")\
5    .load(file_location)
6
7  df.show()
```

> org.apache.spark.SparkException: Job aborted due to stage failure: Task 0 in stage 23.0 failed 1 times, most recent failure: Lost task 0.0 in stage 23.0 (TID 23) (ip-10-172-190-118.us-west-2.compute.internal executor driver): com.databricks.sql.io.FileReadException: Error while reading file dbfs:/FileStore/tables/employees.csv.

10

```
# PRINT CORRUPTED RECORD
emp_schema = StructType(
    [
        StructField("id", StringType(), True),
        StructField("name", StringType(), True),
        StructField("age", IntegerType(), True),
        StructField("salary", IntegerType(), True),
        StructField("address", StringType(), True),
        StructField("nominee", StringType(), True),
        StructField("_corrupt_record", StringType(), True)
    ]
)
```

11

```
df = spark.read.format("csv") \
  .option("inferSchema", "true") \
  .option("header", "true") \
  .option("mode", "PERMISSIVE")\
  .schema(emp_schema)\
  .load(file_location)

df.show(truncate=False)
```

df: pyspark.sql.dataframe.DataFrame = [id: string, name: string ... 5 more fields]

```
+---+--------+---+------+------------+--------+-----------------------------------------+
|id |name    |age|salary|address     |nominee |_corrupt_record                          |
+---+--------+---+------+------------+--------+-----------------------------------------+
|1  |Manish  |26 |75000 |bihar       |nominee1|null                                     |
|2  |Nikita  |23 |100000|uttarpradesh|nominee2|null                                     |
|3  |Pritam  |22 |150000|Bangalore   |India   |3,Pritam,22,150000,Bangalore,India,nominee3|
|4  |Prantosh|17 |200000|Kolkata     |India   |4,Prantosh,17,200000,Kolkata,India,nominee4|
|5  |Vikash  |31 |300000|null        |nominee5|null                                     |
+---+--------+---+------+------------+--------+-----------------------------------------+
```

12

```
df = spark.read.format("csv") \
  .option("inferSchema", "true") \
  .option("header", "true") \
  .schema(emp_schema)\
  .option("badRecordsPath", "/FileStore/tables/bad_records")\
  .load(file_location)

df.show(truncate=False)
```

df: pyspark.sql.dataframe.DataFrame = [id: string, name: string ... 5 more fields]

```
+---+------+---+------+------------+--------+---------------+
|id |name  |age|salary|address     |nominee |_corrupt_record|
```

```
+---+------+---+------+------------+--------+---------------+
|1  |Manish|26 |75000 |bihar       |nominee1|null           |
|2  |Nikita|23 |100000|uttarpradesh|nominee2|null           |
|5  |Vikash|31 |300000|null        |nominee5|null           |
+---+------+---+------+------------+--------+---------------+
```

13

```
%fs
ls /FileStore/tables/bad_records/20241212T024920/bad_records/
```

Table

	path	name
1	dbfs:/FileStore/tables/bad_records/20241212T024920/bad_records/part-00000-b95df98a-9cec-448e-af79-386453539f13	part-00000-b95df98a-9c

1 row

14

```
bad_rec_df = spark.read.format("json").load("/FileStore/tables/bad_records/20241212T024920/bad_records/")

bad_rec_df.show()
```

bad_rec_df: pyspark.sql.dataframe.DataFrame = [path: string, reason: string ... 1 more field]

```
+--------------------+--------------------+--------------------+
|                path|              reason|              record|
+--------------------+--------------------+--------------------+
|dbfs:/FileStore/t...|org.apache.spark....|3,Pritam,22,15000...|
|dbfs:/FileStore/t...|org.apache.spark....|4,Prantosh,17,200...|
+--------------------+--------------------+--------------------+
```

15