

## Read Data in SPARK

1. The core structure of code would be

Core Structure

```
dataframeReader.format( --- ) \
    .option( "key", "value" ) \
    .schema( --- ) \
    .load( - - )
```

2. Here format is data file format(CSV, JSON, ODBC/JSON etc, PARQUET)..in spark the default file format is parquet

format ⇒ Data file format  
CSV, JSON, JDBC/ODBC, Table, parquet  
↳ optional

3. Here in option we have things like

option ⇒ inferSchema, mode, header  
↳ optional

4. In Schema ..we can pass manual schema if we have one
5. And in load we have to give location of our data

Schema ⇒ manual schema you can pass  
↳ optional  
load ⇒ path where our data is residing

6. How to access dataframe reader API? ..for that we have to initialize the spark session

Dataframe Reader API  
↳ how to access ?  
↳ spark.read

and use spark.read

7. Example code would look like this

example

```
spark.read.format("CSV") \
  .option("header", "true") \
  .option("inferSchema", "true") \
  .option("mode", "FAILFAST") \
  .load("c:\user\download\data.csv")
```

Here we didn't use .schema because we have used inferSchema

8. Lets learn about mode

Mode

- ① Fail fast → fail execution if malformed record in dataset.
- ② Drop malformed → Drop the corrupted record
- ③ Permissive → default  
→ set null value to all corrupted fields

9. We have 3 modes and Permissive mode is default mode..it sets null value to all corrupted fields
10. Now we'll do hands-on

```
flight_df_header_schema = spark.read.format("csv")\  
    .option("header", "true")\  
    .option("inferSchema", "true")\  
    .option("mode", "FAILFAST")\  
    .load("/FileStore/tables/flight_data.csv")  
  
flight_df_header.show(5)
```

- 11.
12. Notes on inferSchema : <https://g.co/gemini/share/43be33546bfa>
- 13.