Apache Spark

Spark Data Source API

Spark Data Source API

DataFrameReader API : https://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.sql.DataFrameReader

General structure

```
DataFrameReader
    .format(...)
    .option("key", "value")
    .schema(...)
    .load()
```

Indicative Example

```
spark.read
    .format("csv")
    .option)"header", "true")
    .option("path", "/data/mycsvfiles/")
    .option("mode", "FAILFAST")
    .schema(mySchema)
    .load()
```

Spark Data Source API

DataFrameReader API : https://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.sql.DataFrameReader

General structure

```
DataFrameReader
    .format(...)
    .option("key", "value")
    .schema(...)
    .load()
```

Indicative Example

Built In Formats
CSV, JSON, Parquet, ORC, JDBC

```
spark.read
    .format("csv")
    .option)"header", "true")
    .option("path", "/data/mycsvfiles/")
    .option("mode", "FAILFAST")
    .schema(mySchema)
    .load()
```

Spark Data Source API

DataFrameReader API : https://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.sql.DataFrameReader

General structure

```
DataFrameReader
    .format(...)
    .option("key", "value")
    .schema(...)
    .load()
```

Indicative Example

```
spark.read
    .format("csv")
    .option)"header", "true")
    .option("path", "/data/mycsvfiles/")
    .option("mode", "FAILFAST")
    .schema(mySchema)
    .load()
```

Community Formats
Cassandra, MongoDB, AVRO, XML, HBase, Redshift

Spark Data Source API

DataFrameReader API : https://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.sql.DataFrameReader

General structure

```
DataFrameReader
    .format(...)
    .option("key", "value")
    .schema(...)
    .load()
```

Indicative Example

```
spark.read
    .format("csv")
    .option)"header", "true")
    .option("path", "/data/mycsvfiles/")
    .option("mode", "FAILFAST")
    .schema(mySchema)
    .load()
```

Read Mode
1. PERMISSIVE
2. DROPMALFORMED
3. FAILFAST

Spark Data Source API

DataFrameReader API : https://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.sql.DataFrameReader

General structure

```
DataFrameReader
    .format(...)
    .option("key", "value")
    .schema(...)
    .load()
```

Indicative Example

```
spark.read
    .format("csv")
    .option)"header", "true")
    .option("path", "/data/mycsvfiles/")
    .option("mode", "FAILFAST")
    .schema(mySchema)
    .load()
```

Schema
1. Explicit
2. Infer Schema
3. Implicit


LEARNING JOURNAL
www.learningjournal.guru