1. Сделал свою схему и загрузил датасет billionaires.csv с kaggle

```
Using Python version 2.7.5 (default, Apr  2 2020 13:16:51)
SparkSession available as 'spark'.
>>> from pyspark.sql import SparkSession
>>> from pyspark.sql import functions as F
>>> from pyspark.sql.types import StructType, StringType, IntegerType, FloatType
>>> schema = StructType() \
...     .add("year", IntegerType()) \
...     .add("rank", IntegerType()) \
...     .add("name", StringType()) \
...     .add("net_worth", FloatType()) \
...     .add("age", IntegerType()) \
...     .add("natinality", StringType()) \
...     .add("source_wealth", StringType())
>>> raw_files = spark \
...     .readStream \
...     .format("csv") \
...     .schema(schema) \
...     .options(path="csv_stream", header=True) \
...     .load()
```

```
-------------------------------------------
+----+----+-----------------+---------+----+-------------+--------------------------+
|year|rank|name             |net_worth|age |natinality   |source_wealth             |
+----+----+-----------------+---------+----+-------------+--------------------------+
|null|null|null             |null     |null|null         |null                      |
|2019|2   |Bill Gates       |96.5     |63  |United States|Microsoft                 |
|2019|3   |Warren Buffett   |82.5     |88  |United States|Berkshire Hathaway        |
|null|null|null             |null     |null|null         |null                      |
|null|null|null             |null     |null|null         |null                      |
|2019|6   |Amancio Ortega   |62.7     |82  |Spain        |Inditex, Zara             |
|2019|7   |Larry Ellison    |62.5     |74  |United States|Oracle Corporation        |
|2019|8   |Mark Zuckerberg  |62.3     |34  |United States|Facebook                  |
|2019|9   |Michael Bloomberg|55.5     |77  |United States|Bloomberg L.P.            |
|2019|10  |Larry Page       |50.8     |45  |United States|Alphabet Inc.             |
|null|null|null             |null     |null|null         |null                      |
|null|null|null             |null     |null|null         |null                      |
|null|null|null             |null     |null|null         |null                      |
|null|null|null             |null     |null|null         |null                      |
|2018|5   |Mark Zuckerberg  |71.0     |33  |United States|Facebook                  |
|null|null|null             |null     |null|null         |null                      |
|2018|7   |Carlos Slim      |67.1     |78  |Mexico       |América Móvil, Grupo Carso|
|null|null|null             |null     |null|null         |null                      |
|null|null|null             |null     |null|null         |null                      |
|2018|10  |Larry Ellison    |58.5     |73  |United States|Oracle Corporation        |
+----+----+-----------------+---------+----+-------------+--------------------------+
only showing top 20 rows
```

2. Далее сделал топик irisTopic и загрузил туда json файл с датасетом ирис с kaggle. Прочитал сырые данные и потом спарсил и прочитал форматированные данные.

```
>>> raw_iris = spark.read. \
...     format("kafka"). \
...     option("kafka.bootstrap.servers", kafka_brokers). \
...     option("subscribe", "irisTopic"). \
...     option("startingOffsets", "earliest"). \
...     load()
>>>
>>> raw_iris.show()
+----+--------------------+---------+---------+------+--------------------+-------------+
| key|               value|    topic|partition|offset|           timestamp|timestampType|
+----+--------------------+---------+---------+------+--------------------+-------------+
|null|[20 20 7B 22 73 6...|irisTopic|        0|     0|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|     1|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|     2|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|     3|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|     4|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|     5|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|     6|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|     7|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|     8|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|     9|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|    10|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|    11|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|    12|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|    13|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|    14|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|    15|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|    16|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|    17|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|    18|2020-12-16 02:42:...|            0|
|null|[20 20 7B 22 73 6...|irisTopic|        0|    19|2020-12-16 02:42:...|            0|
+----+--------------------+---------+---------+------+--------------------+-------------+
only showing top 20 rows
```

```
>>> schema = StructType() \
...     .add("sepalLength", FloatType()) \
...     .add("sepalWidth", FloatType()) \
...     .add("petalLength", FloatType()) \
...     .add("petalWidth", FloatType()) \
...     .add("species", StringType())
>>>
>>> value_iris = raw_iris \
...     .select(F.from_json(F.col("value").cast("String"), schema).alias("value"), "offset")
>>>
>>> value_iris.printSchema()
root
 |-- value: struct (nullable = true)
 |    |-- sepalLength: float (nullable = true)
 |    |-- sepalWidth: float (nullable = true)
 |    |-- petalLength: float (nullable = true)
 |    |-- petalWidth: float (nullable = true)
 |    |-- species: string (nullable = true)
 |-- offset: long (nullable = true)

>>>
>>> parsed_iris = value_iris.select("value.*", "offset")
>>>
>>> parsed_iris.printSchema()
root
 |-- sepalLength: float (nullable = true)
 |-- sepalWidth: float (nullable = true)
 |-- petalLength: float (nullable = true)
 |-- petalWidth: float (nullable = true)
 |-- species: string (nullable = true)
 |-- offset: long (nullable = true)

>>> parsed_iris.show()
+-----------+----------+-----------+----------+----------+------+
|sepalLength|sepalWidth|petalLength|petalWidth|   species|offset|
+-----------+----------+-----------+----------+----------+------+
|        4.9|       3.0|        1.4|       0.2|    setosa|     0|
|        5.0|       3.6|        1.4|       0.2|    setosa|     1|
|        5.0|       3.4|        1.5|       0.2|    setosa|     2|
|        5.4|       3.7|        1.5|       0.2|    setosa|     3|
|        4.3|       3.0|        1.1|       0.1|    setosa|     4|
|        5.4|       3.9|        1.3|       0.4|    setosa|     5|
|        5.1|       3.8|        1.5|       0.3|    setosa|     6|
|        4.6|       3.6|        1.0|       0.2|    setosa|     7|
|        5.0|       3.0|        1.6|       0.2|    setosa|     8|
|        5.2|       3.4|        1.4|       0.2|    setosa|     9|
|        5.4|       3.4|        1.5|       0.4|    setosa|    10|
|        4.9|       3.1|        1.5|       0.2|    setosa|    11|
|        4.9|       3.6|        1.4|       0.1|    setosa|    12|
|        5.0|       3.5|        1.3|       0.3|    setosa|    13|
|        5.0|       3.5|        1.6|       0.6|    setosa|    14|
|        5.1|       3.8|        1.6|       0.2|    setosa|    15|
|        5.0|       3.3|        1.4|       0.2|    setosa|    16|
|        6.9|       3.1|        4.9|       1.5|versicolor|    17|
|        5.7|       2.8|        4.5|       1.3|versicolor|    18|
|        6.6|       2.9|        4.6|       1.3|versicolor|    19|
+-----------+----------+-----------+----------+----------+------+
only showing top 20 rows
```