DataOps.onETL

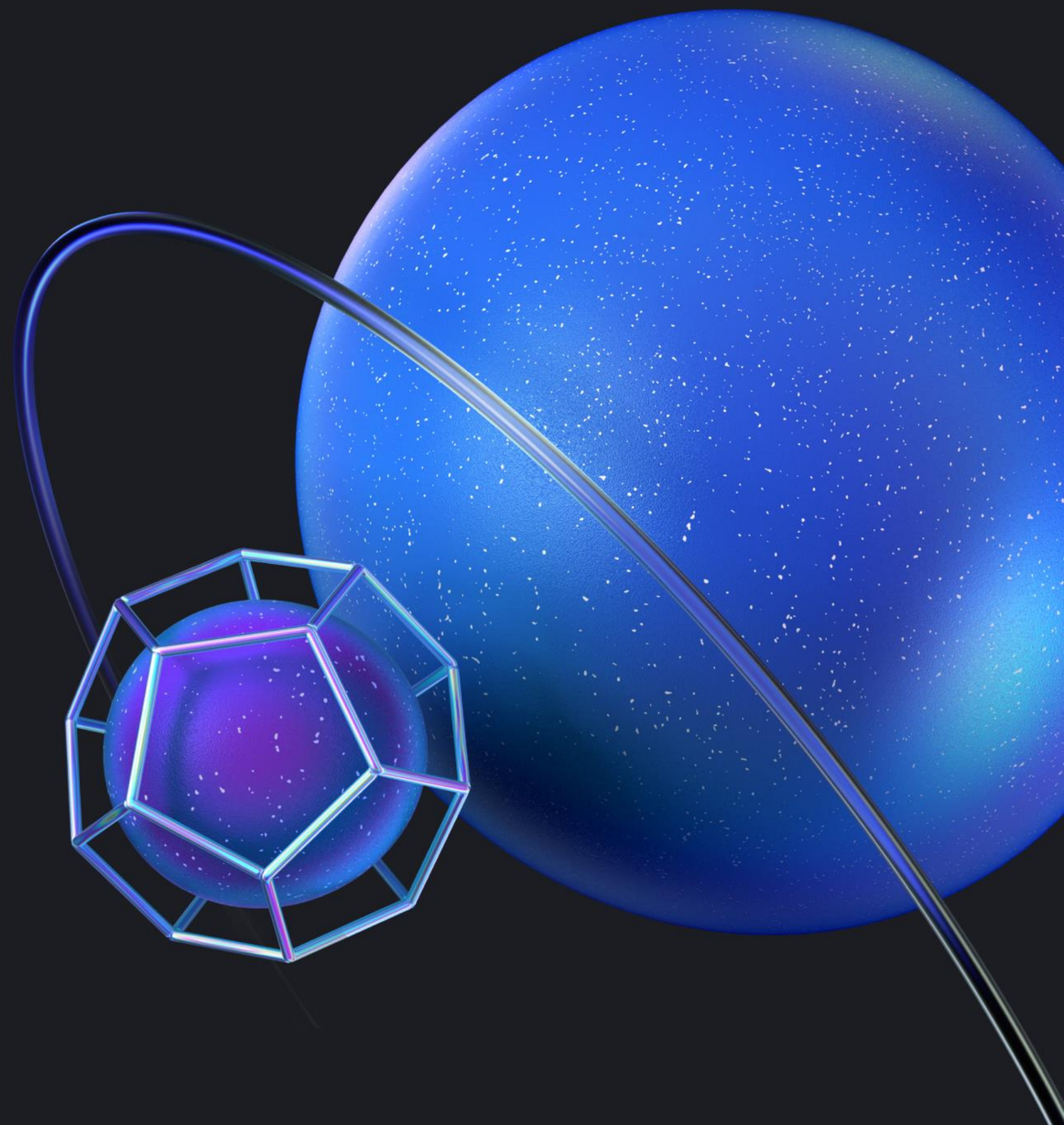
Unit#3

Объекты манипуляции данными DBReader и DBWriter

Саттар Гюльмамедов

PO команды ETL

МТС Тета x DataOps Platform



onETL урок # 3

→ DBReader

→ DBWriter

→ Назначение и отличие от коннекторов

→ Опции конструкторов и методы

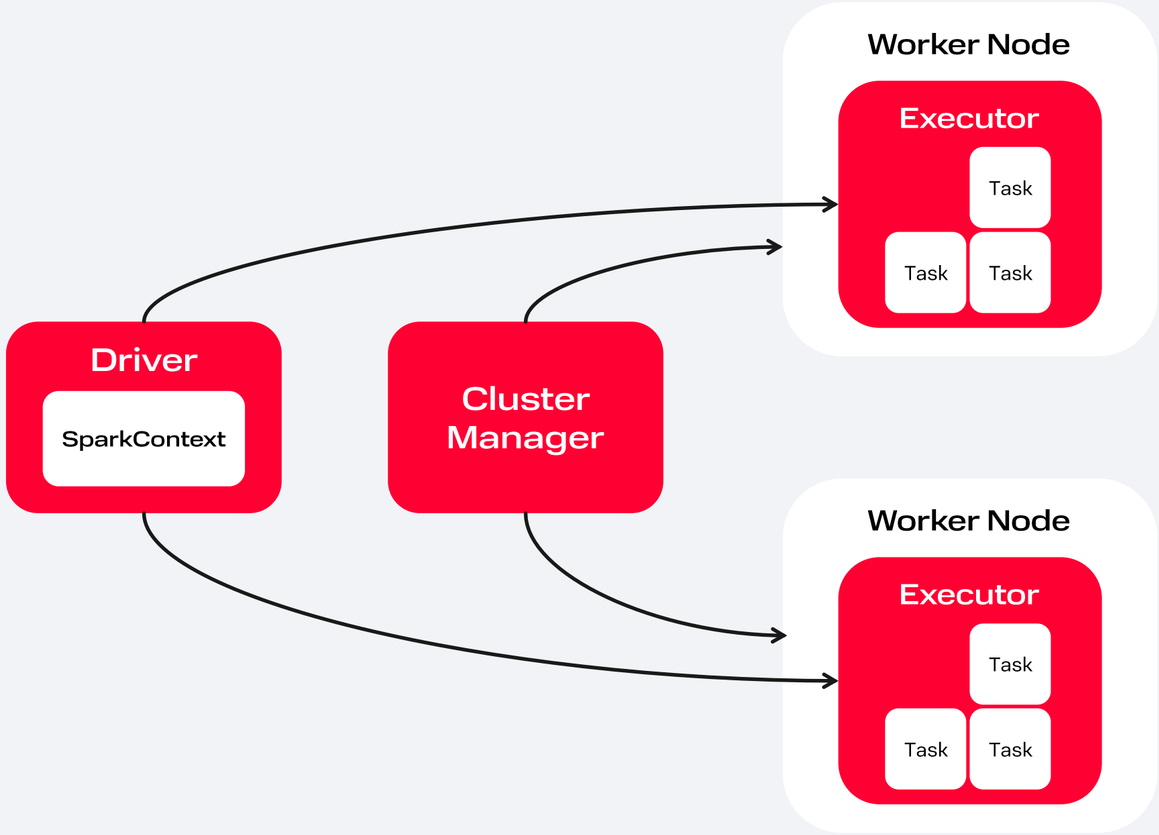
→ Где найти документацию

Отличия от коннекторов

Функциональность	Коннекторы	Объекты чтения/записи
Установка соединения	+	-
Методы для driver	+	-
Методы для executor	+	+
Использование стратегий	-	+
Модуль	onetl.connection	onetl.db

Отличия от коннекторов

https://spark.apache.org/docs/latest/cluster-overview.html



DBReader

→ данные в Spark DataFrame

→ метод run()

https://onetl.readthedocs.io/en/stable/db/db_reader.html

DBReader() - нюансы

Важно	Минусы	Плюсы
Read Strategy Matters	Одна таблица	HWM
		df_schema

DBReader() - параметры

→ connection

→ source (alias “table”)

DBReader() - columns

Python

```python
columns = [
    "mycolumn",
    "another_column as alias",
    "count(*) over ()",
    "some(function) as alias2",
]
```

DBReader() - where

→ MySQL

```python
where = "column_1 > 2"
```

→ MongoDB

```python
where = {
    "col_1": {"$gt": 1, "$lt": 100},
    "col_2": {"$gt": 2},
    "col_3": {"$eq": "hello"},
}
```

DBReader() - hwm

Python

```python
from onetl.hwm import AutoDetectHWM

hwm = AutoDetectHWM(
    name="some_unique_hwm_name",
    expression="hwm_column",
)
```

DBReader() - hint

→ Oracle

```python
hint = "index(myschema.mytable mycolumn)"
```

→ MongoDB

```python
hint = {
    "mycolumn": 1,
}
```

DBReader() - df_schema

Python

```python
from pyspark.sql.types import (
    DoubleType,
    IntegerType,
    StringType,
    StructField,
    StructType,
    TimestampType,
)

df_schema = StructType(
    [
        StructField("_id", IntegerType()),
        StructField("text_string", StringType()),
        StructField("hwm_int", IntegerType()),
        StructField("hwm_datetime",
TimestampType()),
        StructField("float_value", DoubleType()),
    ],
)
```

DBReader() - options

https://spark.apache.org/docs/latest/sql-data-sources-jdbc.html

Опции чтения

option	Умолчание	Возможные значения
partitioning_mode	range	range, hash, mod
partition_column	умолчания нет	
num_partitions	1	
lower_bound	умолчания нет	
upper_bound	умолчания нет	
session_init_statement	умолчания нет	
query_timeout	может быть установлено jdbc-драйвером	
fetchsize	10 000	

run()

Python

```python
from onetl.db import DBReader
from onetl.connection import Hive
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("spark-app-name") \
    .enableHiveSupport() \
    .getOrCreate()

hive = Hive(cluster="rnd-dwh", spark=spark)

reader = DBReader(connection=hive, source="fiddle.dummy")
df = reader.run()
```

DBWriter

→ конструктор

→ метод run(df)

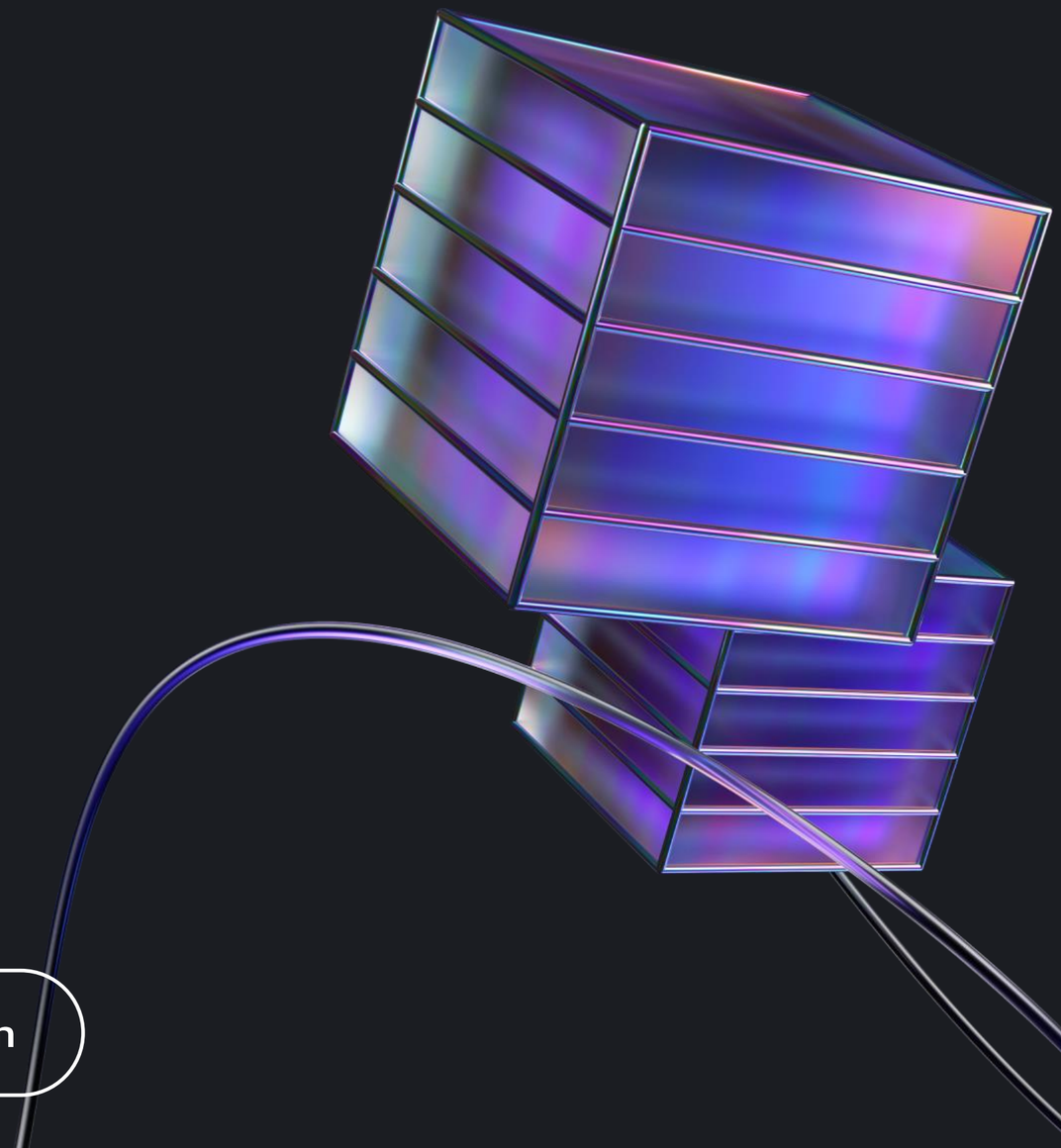
https://onetl.readthedocs.io/en/stable/db/db_writer.html

DBReader() - параметры

→ connection

→ target (alias “table”)

DBReader() – options



МТС Тета x DataOps Platform

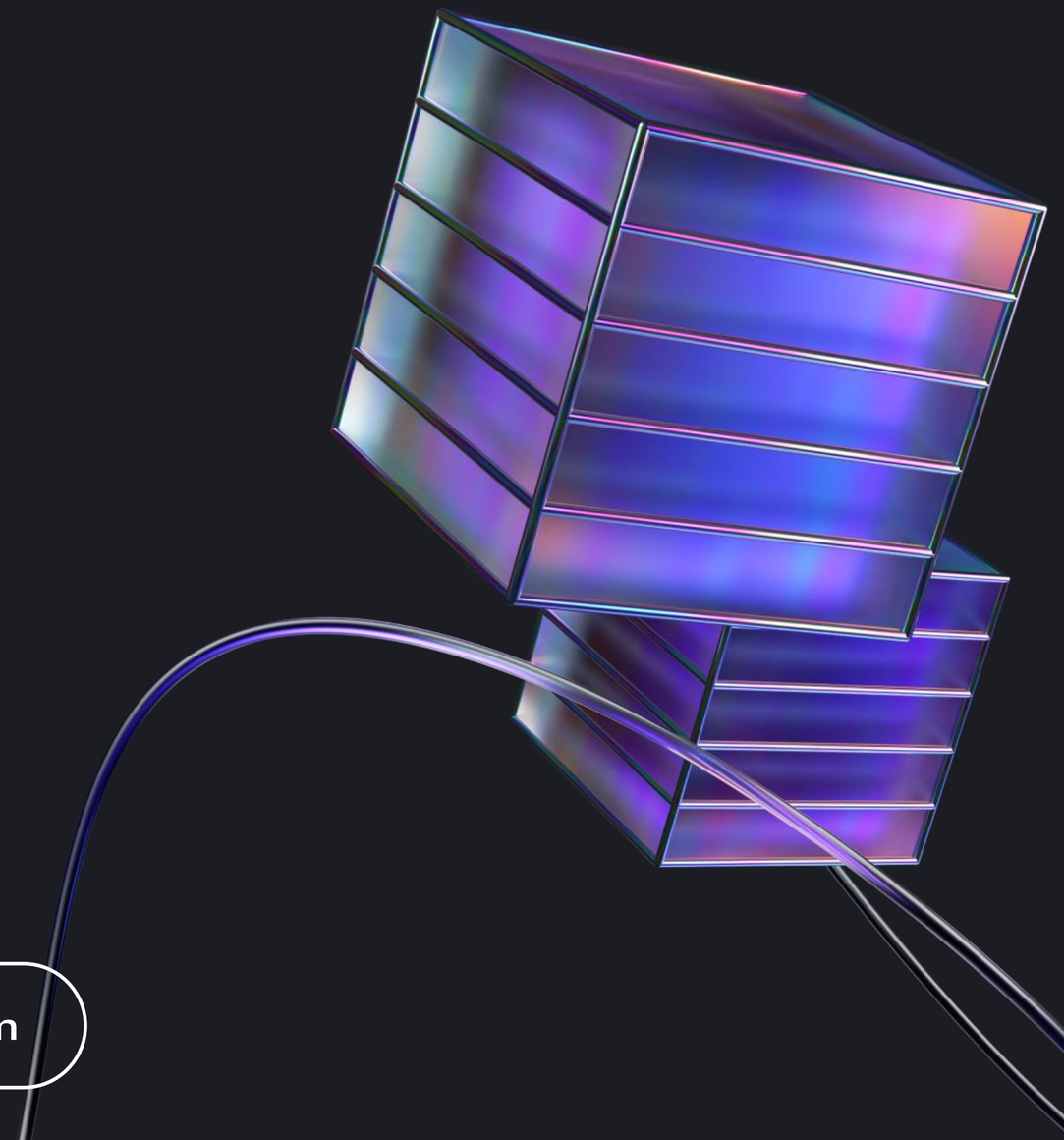
DBReader() - options

option	Умолчание	Возможные значения
query_timeout	может быть установлено jdbc-драйвером	
fetchsize	может быть установлено jdbc-драйвером	
if_exists	append	append, replace_entire_table, ignore, error
batchsize	20 000	
isolation_level	READ_UNCOMMITTED	NONE, READ_COMMITTED, READ_UNCOMMITTED, REPEATABLE_READ, SERIALIZABLE

run(df)

```python
from onetl.db import DBReader, DBWriter
from onetl.connection import Hive
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("spark-app-
name").enableHiveSupport().getOrCreate()

hive = Hive(cluster="rnd-dwh", spark=spark)

reader = DBReader(connection=hive, source="fiddle.dummy")
df = reader.run()

options = {"compression": "snappy", "partitionBy": "id"}

writer = DBWriter(
    connection=hive,
    target="default.test",
    options=options,
)

writer.run(df)
```

Демо



МТС Тета x DataOps Platform

onETL урок # 3

→ DBReader

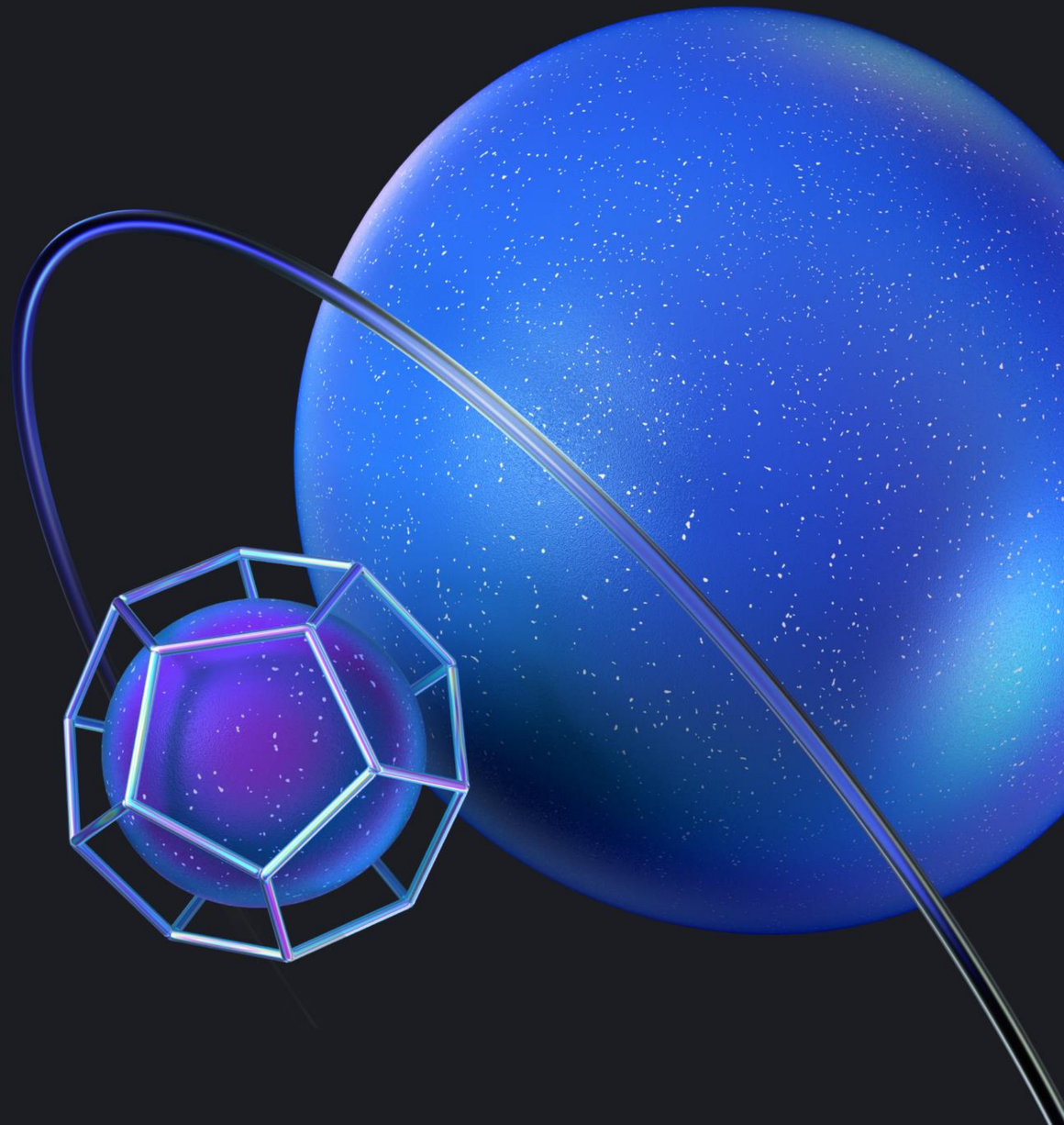
→ DBWriter

→ Назначение и отличие от коннекторов

→ Опции конструкторов и методы

→ Где найти документацию

Спасибо!



onetools@mts.ru
https://t.me/c/1511728757/5

МТС Тета x DataOps Platform