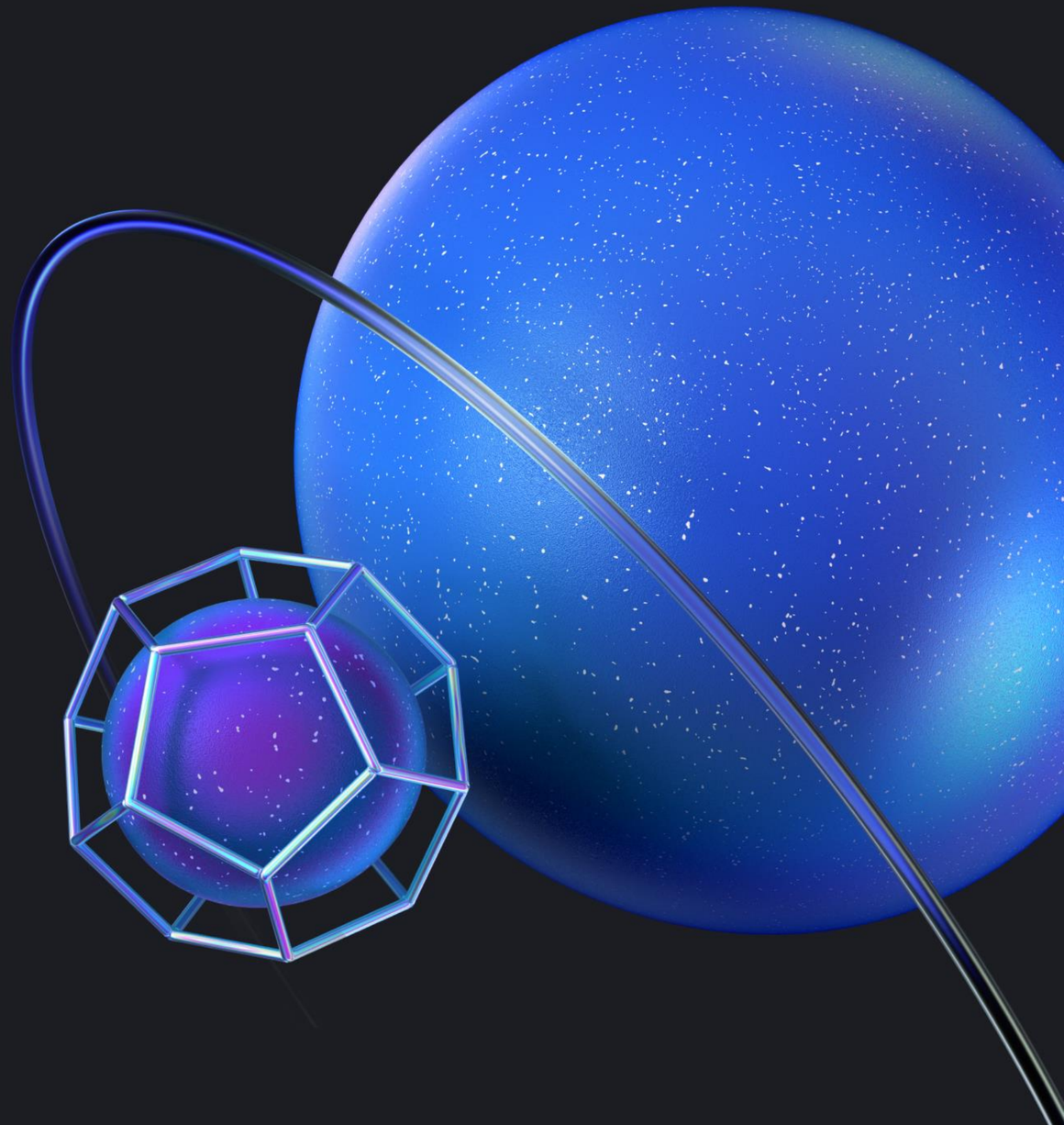
DataOps.onETL

Unit#1

Вводный урок. onETL: что это такое, зачем он нужен, как его получить, варианты использования



Саттар Гюльмамедов

PO команды ETL

МТС Тета x DataOps Platform

onETL урок # 1

→ Назначение

→ Структура

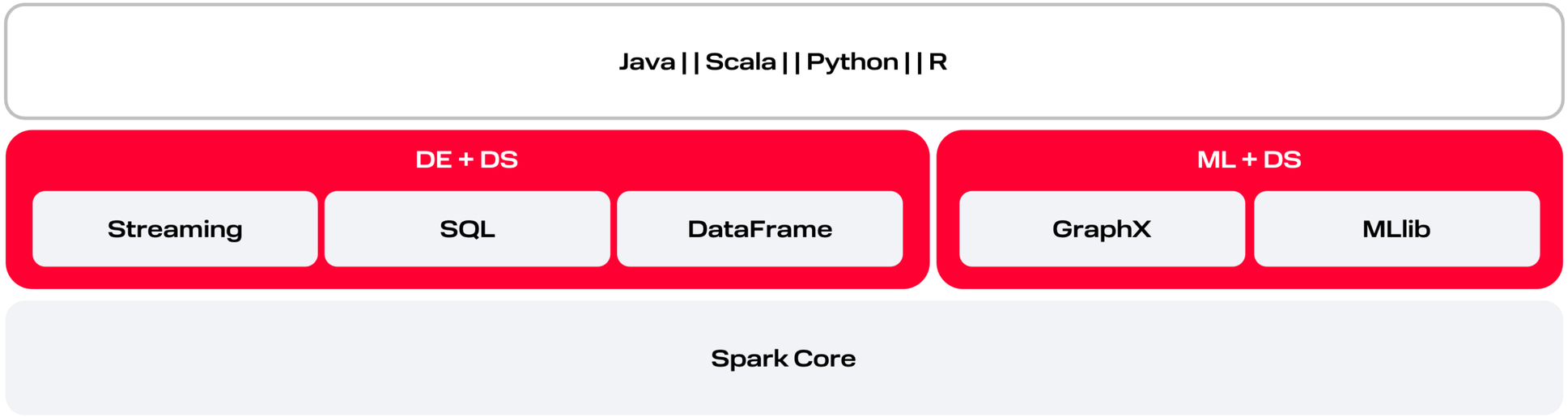
→ Интерфейсы и объекты

→ Как установить

→ Как узнать больше

→ Варианты запуска

Apache Spark



onETL функциональность

→ DBConnections

→ FileConnections, FileDataFrameConnections

→ DBClasses

→ FileClasses

→ FileDataFrameClasses

→ HWM, Read Strategies и HWM Store

DB Connections

Реляционные СУБД

→ Postgres

→ Greenplum

→ Oracle

→ MSSQL

→ MySQL

Нереляционные хранилища

→ Clickhouse

→ Teradata

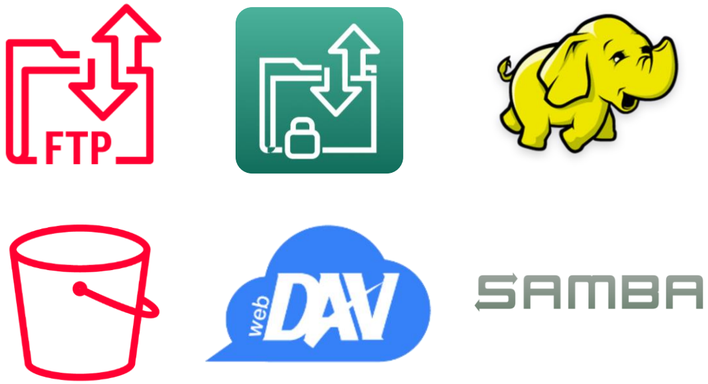
→ Hive

→ Kafka

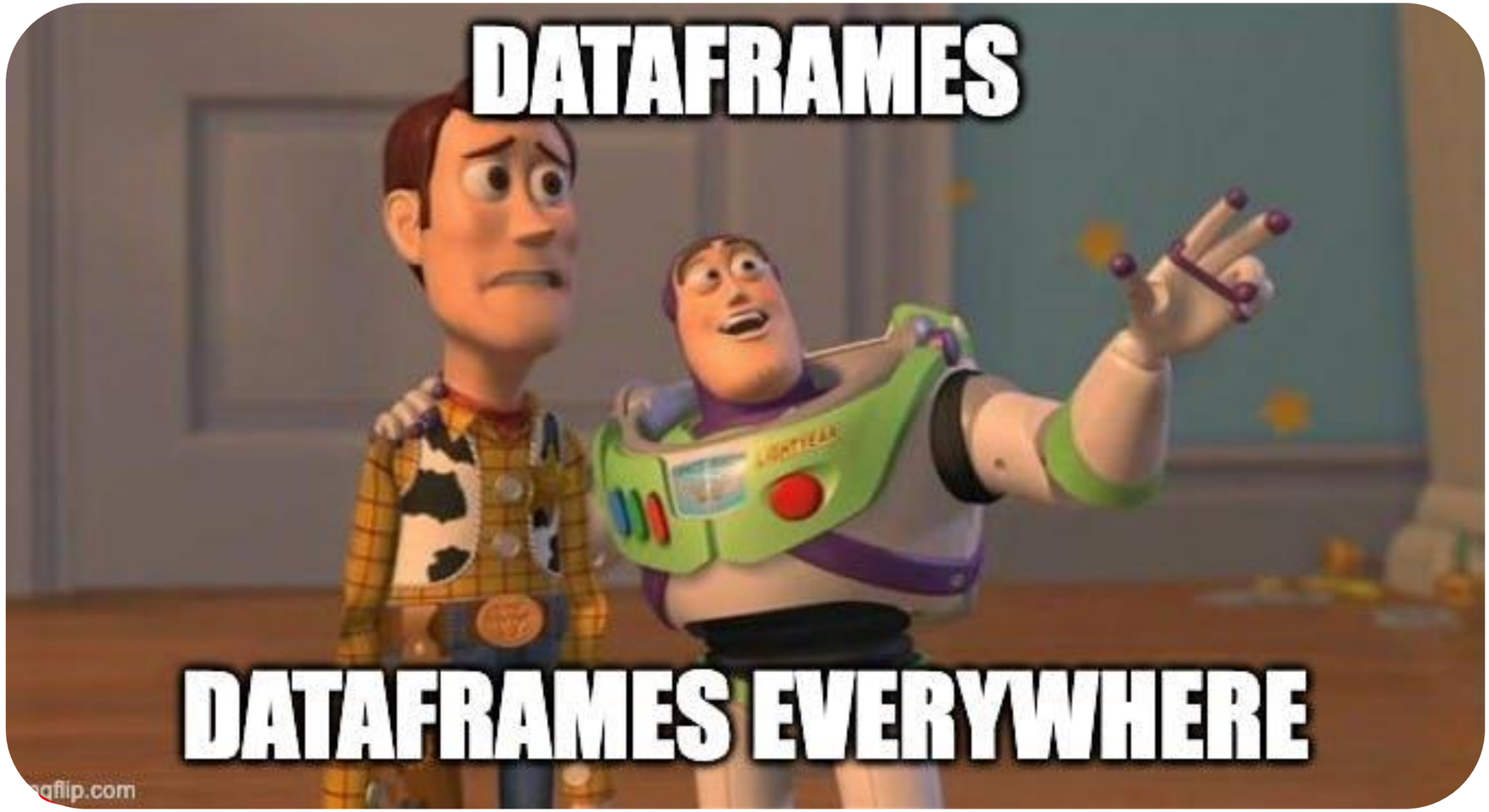
→ MongoDB

FileConnections

- → FTP
- → FTPS
- → SFTP
- → HDFS
- → S3
- → WebDAV
- → Samba



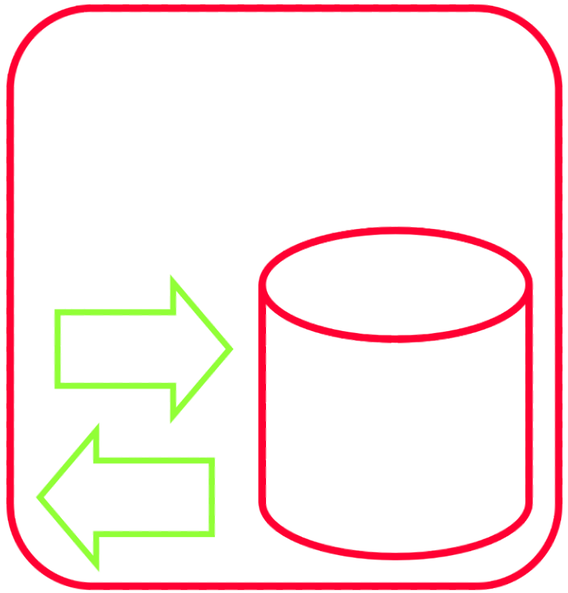
File DataFrame Connections



DBClasses

→ DBReader

→ DBWriter



FileClasses

→ File Downloader

→ File Uploader

→ File Mover

→ Filters

→ Limits

HWM

→ **HWM**

→ Integer

→ Decimal

→ Date

→ Datetime

HWM

- → HWM
- → **Read Strategies**
  - → Snapshot strategy
  - → Incremental strategy
  - → Snapshot batch strategy
  - → Incremental batch strategy

HWM

→ HWM

→ Read Strategies

→ **HWM Store**

→ YAML HWM Store

→ In-memory HWM Store

→ DataOps.ETL Service

https://confluence.mts.ru/display/DataOps/DataOps.ETL

*только для внутренних сотрудников компании МТС

Документация

https://onetl.readthedocs.io

Компетенции

→ python

→ pip

→ Apache Spark

https://spark.apache.org/

Документация Spark

https://spark.apache.org/docs/latest/

Установка в контуре МТС

→ pip.conf

```
[global]
format           = columns
no-cache-dir     = yes
index-url        = https://nexus.services.mts.ru/repository/pip/simple
extra-index-url = https://artifactory.mts.ru/artifactory/api/pypi/own-onetl-pypi-local/simple
trusted-host     = nexus.services.mts.ru
                   artifactory.mts.ru
```

https://pip.pypa.io/en/stable/topics/configuration/#location

Бандлы onETL

→ pip install onetl

→ pip install onetl[spark]

→ pip install onetl[kerberos]

→ pip install onetl[files]

→ pip install onetl[all]

https://onetl.readthedocs.io/en/stable/install/index.html

Поддержка Kerberos

https://onetl.readthedocs.io/en/stable/install/kerberos.html

MTSpark

https://confluence.mts.ru/pages/viewpage.action?pageId=723527060

*Только для внутренних сотрудников компании МТС

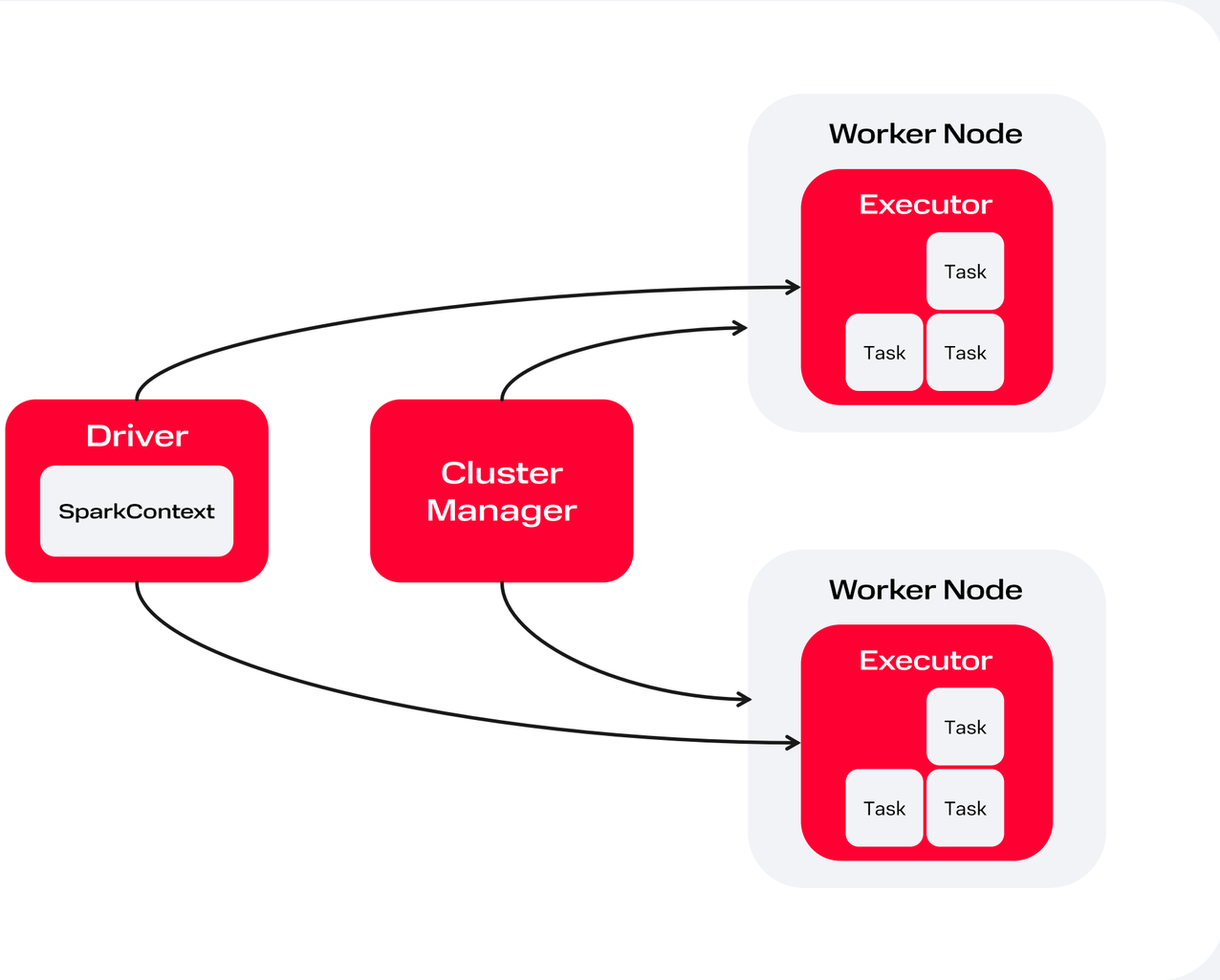
Совместимость со Spark

https://onetl.readthedocs.io/en/stable/install/spark.html#compatibility-matrix

Архитектура Spark

→ Запуск в режиме кластера:
- под управлением Yarn
- под управлением k8s
- Spark Standalone
- Mesos (deprecated)

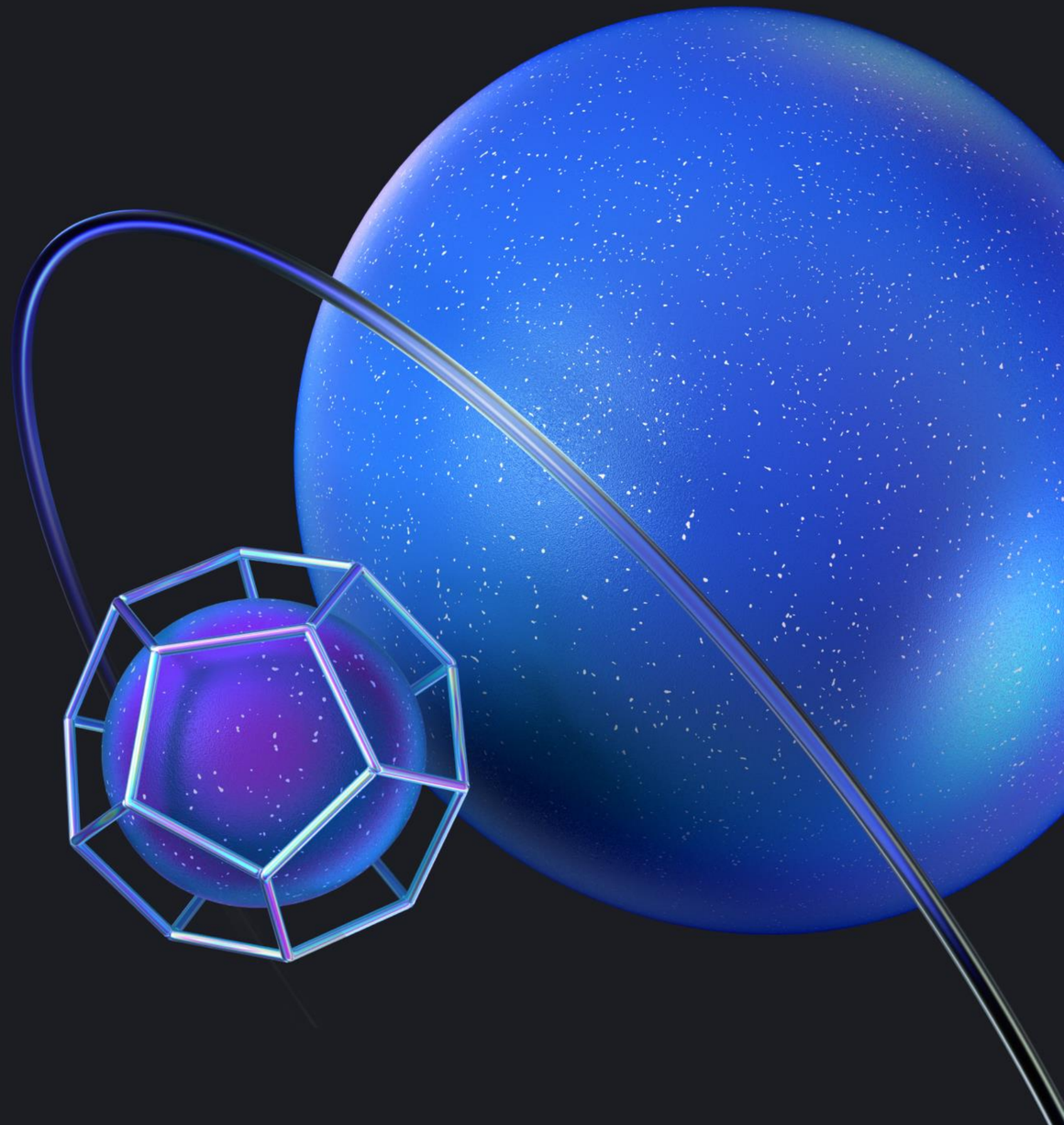
→ Бескластерный Spark



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


SPARK DRIVER
df.collect()

Спасибо!



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

МТС Тета x DataOps Platform