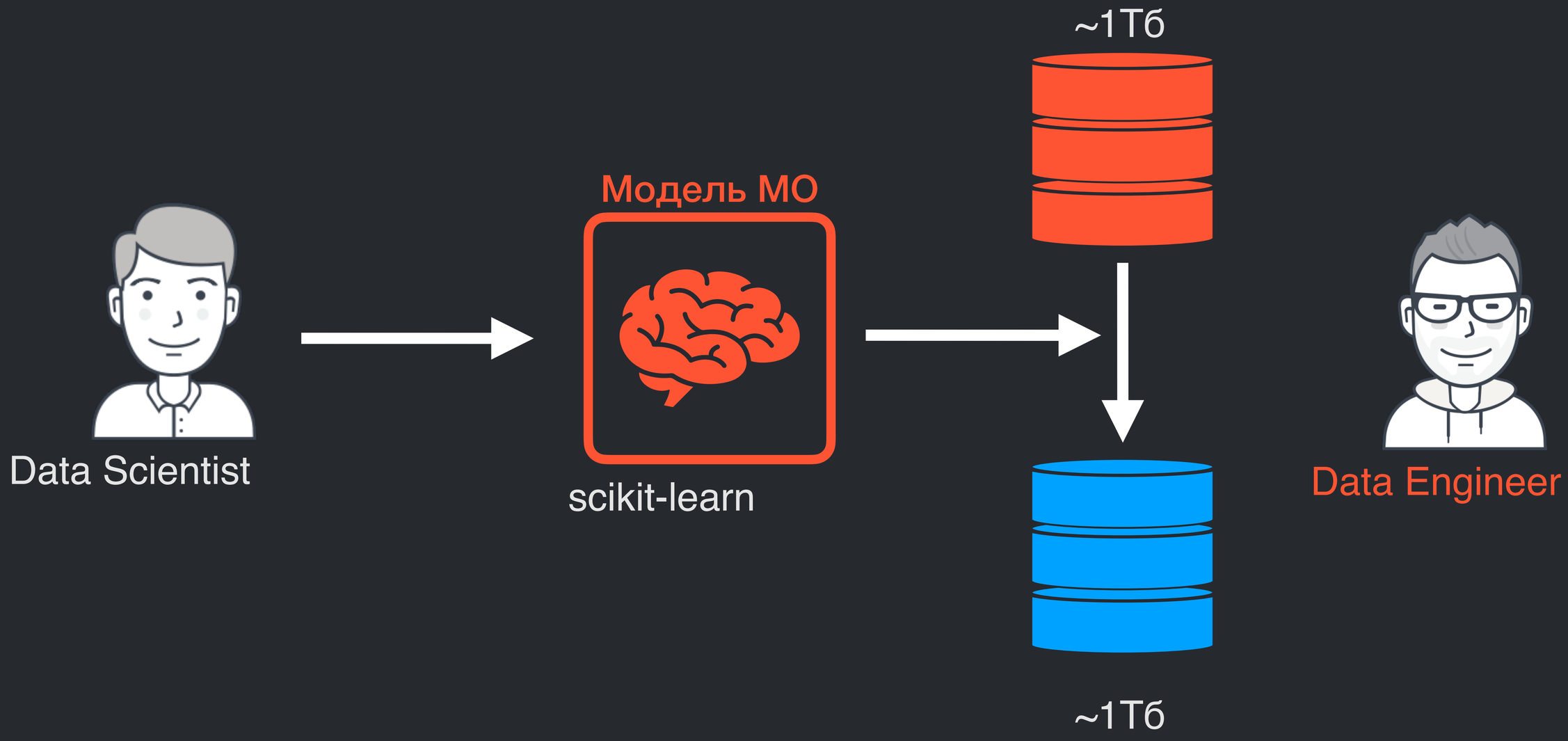
ПРИМЕНЕНИЕ НЕ РАСПРЕДЕЛЕННЫХ МОДЕЛЕЙ НА SPARK

KARPOV.COURSES

BIG DATA & ML


~1Тб
Модель МО
Data Scientist
scikit-learn
Data Engineer
~1Тб

SPARK UDF

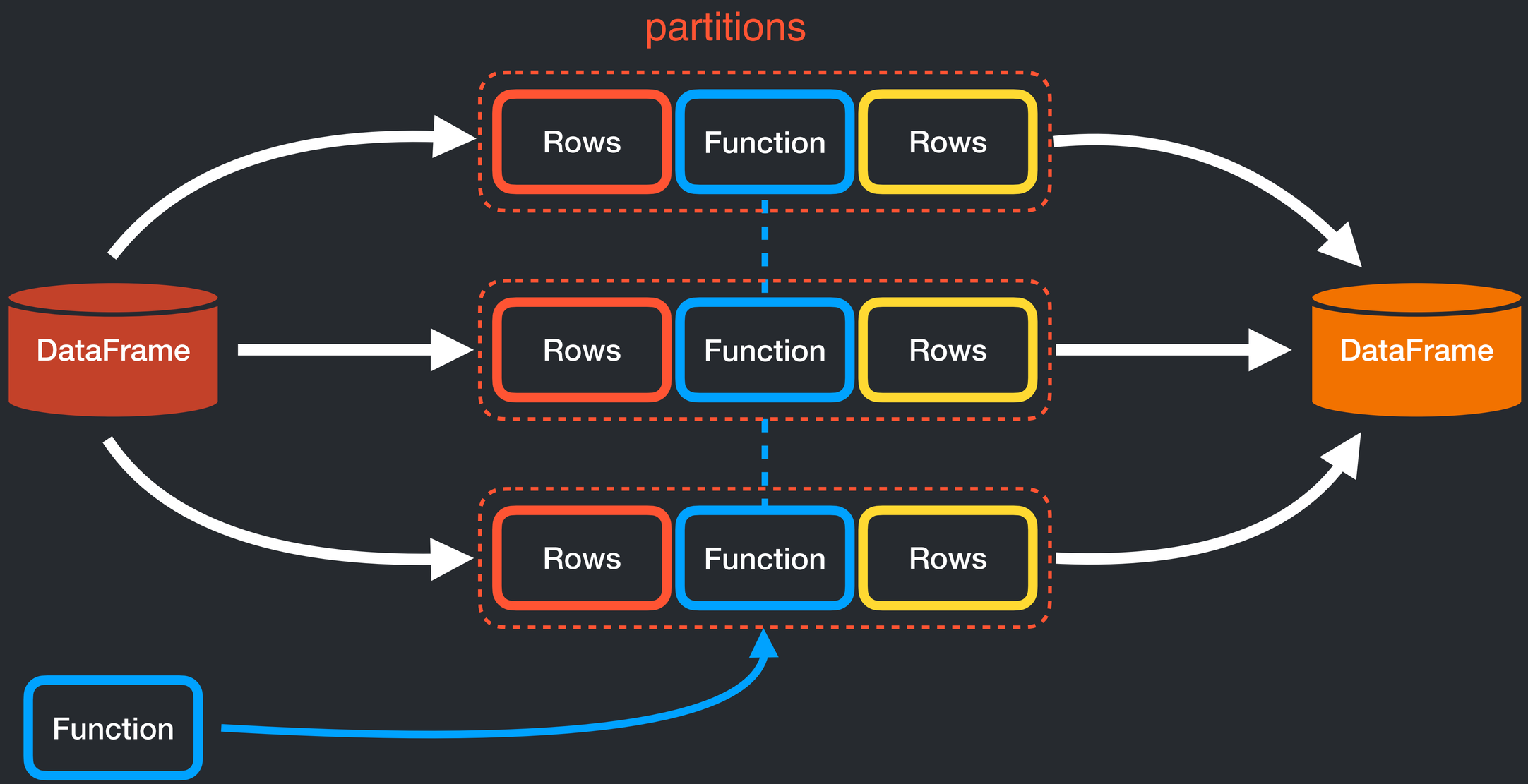
UDF (User Define Function) - это функция Spark, которая позволяет пользователям определять свои собственные функции.

```
from pyspark.sql.functions import udf

@udf('double')
def plus_one(v):
    return v + 1

df.withColumn(‘plus_one’, plus_one(df.value))
```

SPARK UDF


partitions
Rows
Function
Rows
DataFrame
Rows
Function
Rows
DataFrame
Rows
Function
Rows
Function

SPARK PANDAS UDF

Apache Arrow - это столбчатый формат данных (InMemory), который используется в Spark для эффективной передачи данных между процессами JVM и Python.

Pandas UDF (User Define Function) - функции на основе Apache Arrow, предоставляют возможность полностью определять высокопроизводительные пользовательские функции низкими накладными расходами на Python.

SPARK PANDAS UDF


```
from pyspark.sql.functions import pandas_udf

@pandas_udf('double')
def plus_one(v):
    return v + 1

df.withColumn('plus_one', plus_one(df.value))
```

SPARK PANDAS UDF

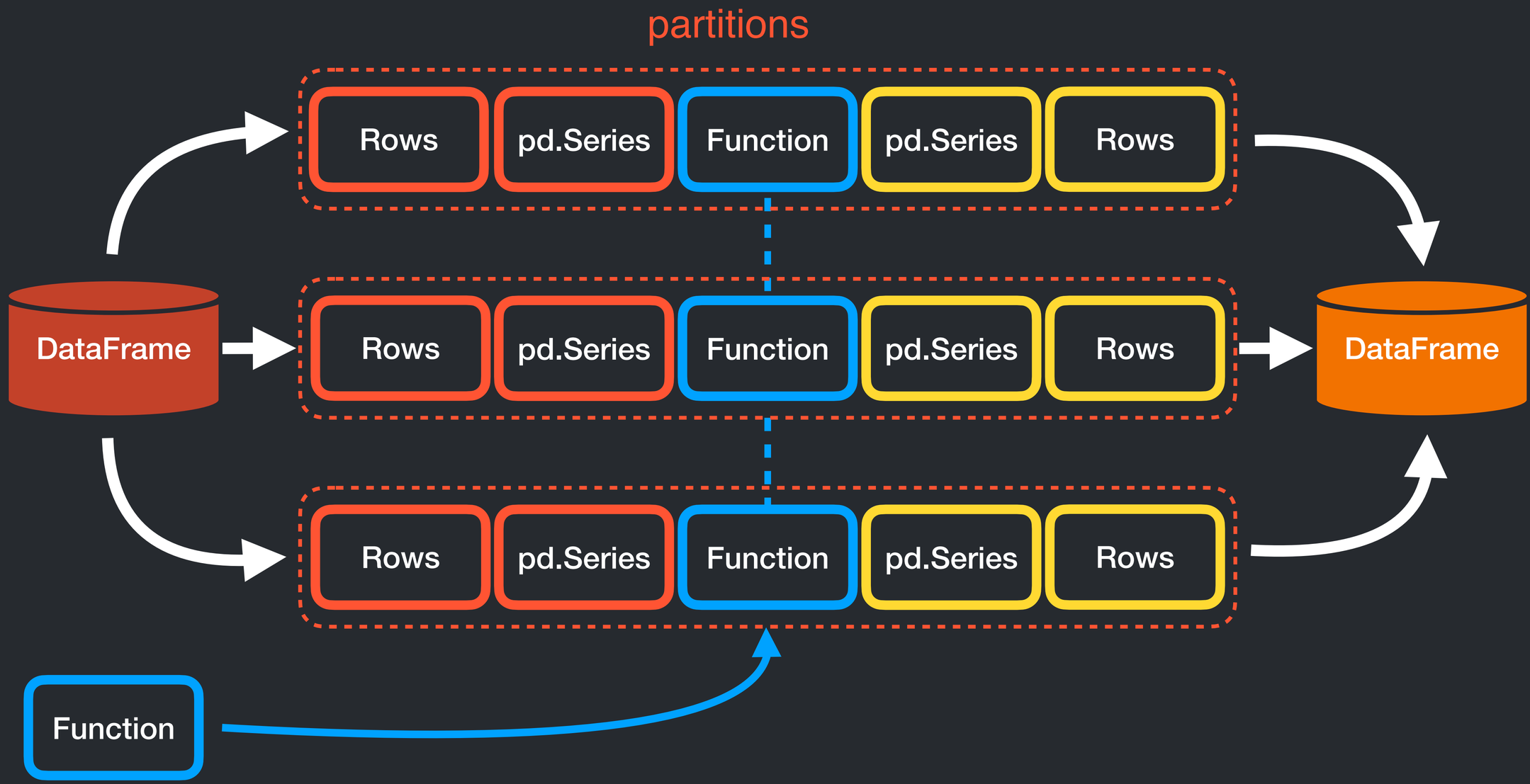

```
from pyspark.sql.functions import pandas_udf, PandasUDFType


@pandas_udf('double', PandasUDFType.SCALAR)
def pudf_mean(v: pd.Series) -> float:
    return v.mean()

df.withColumn('mean', pudf_mean(df.value))
```

SPARK PANDAS UDF

APACHE ARROW

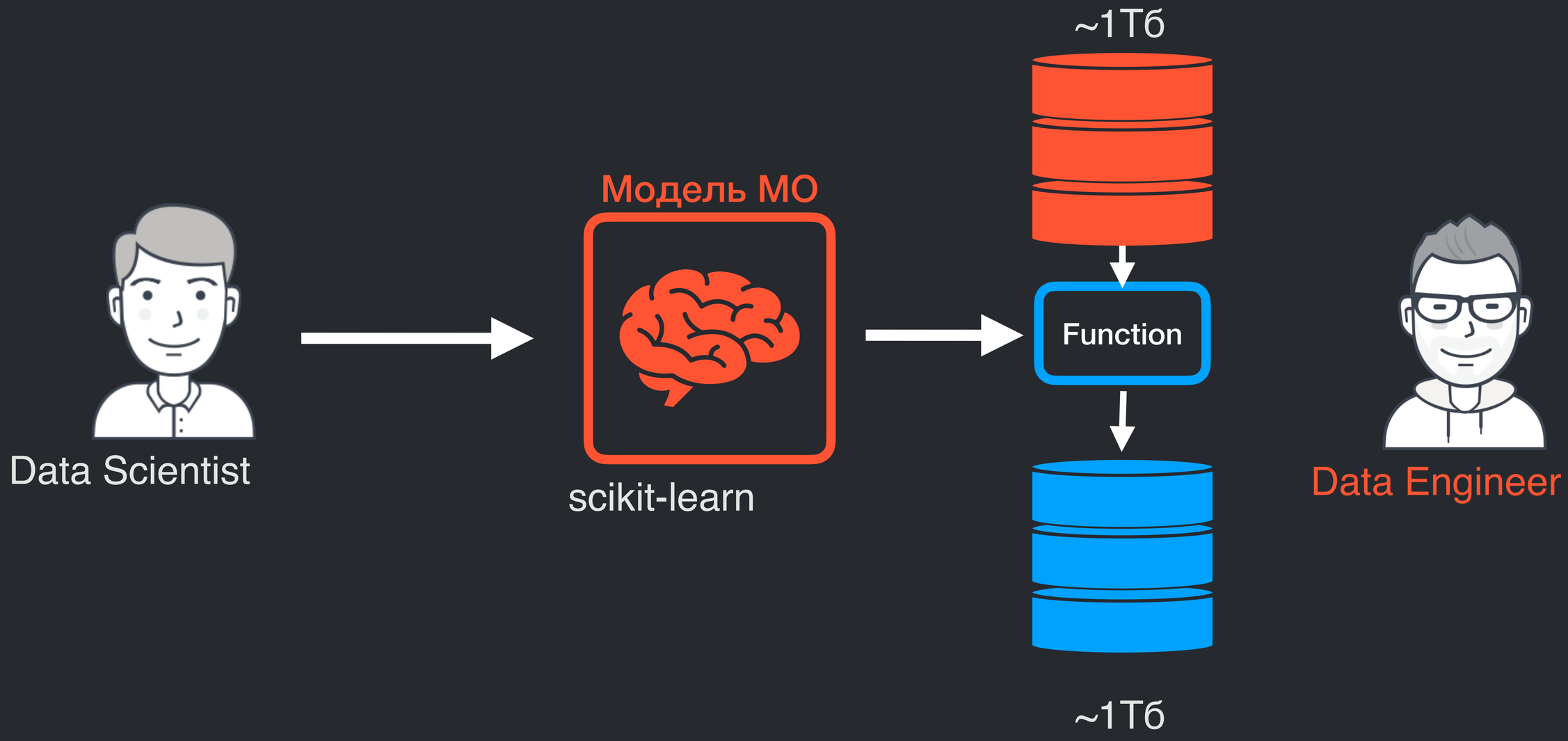


SPARK PANDAS UDF

PandasUDFType

	SCALAR	GROUPED_MAP
Input	pandas.Series	pandas.DataFrame
Output	pandas.Series	pandas.DataFrame
Output size	same as input size	any size
Return Type	DataType (double, int …)	StructType

BIG DATA & ML


~1Тб
Модель МО
Function
Data Scientist
scikit-learn
Data Engineer
~1Тб