Spark Streaming

SELEZNEV ARTEM
HEAD OF CVM ANALYTICS @ MAGNIT

WELCOME STREAMS

STREAMING


Kafka
Flume
HDFS/S3
Kinesis
Twitter
Spark
Streaming
HDFS
Databases
Dashboards

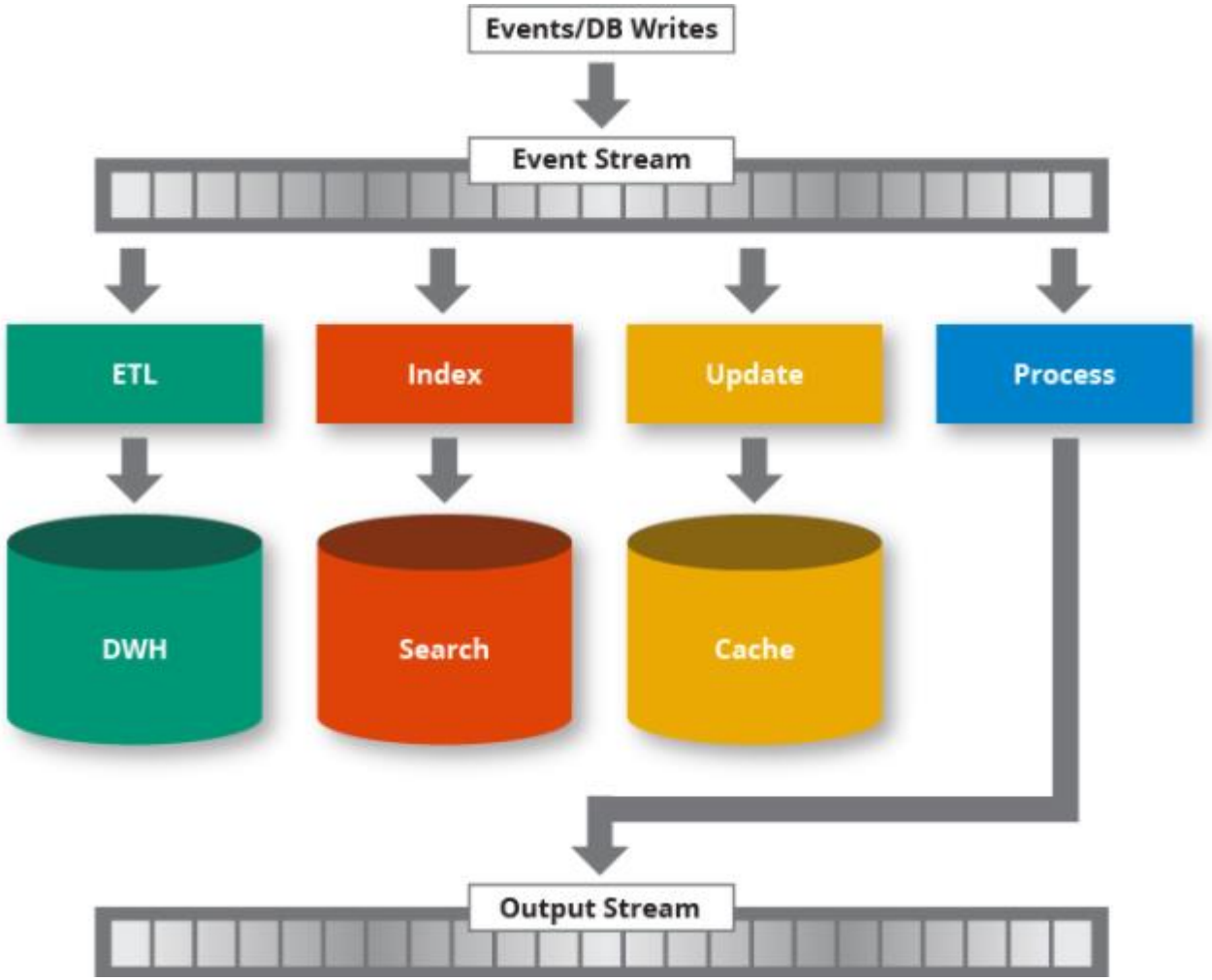
STREAMING


Kafka
Flume
HDFS/S3
Kinesis
Twitter
Spark Streaming
HDFS
Databases
Dashboards


input data stream
Spark Streaming
batches of input data
Spark Engine
batches of processed data

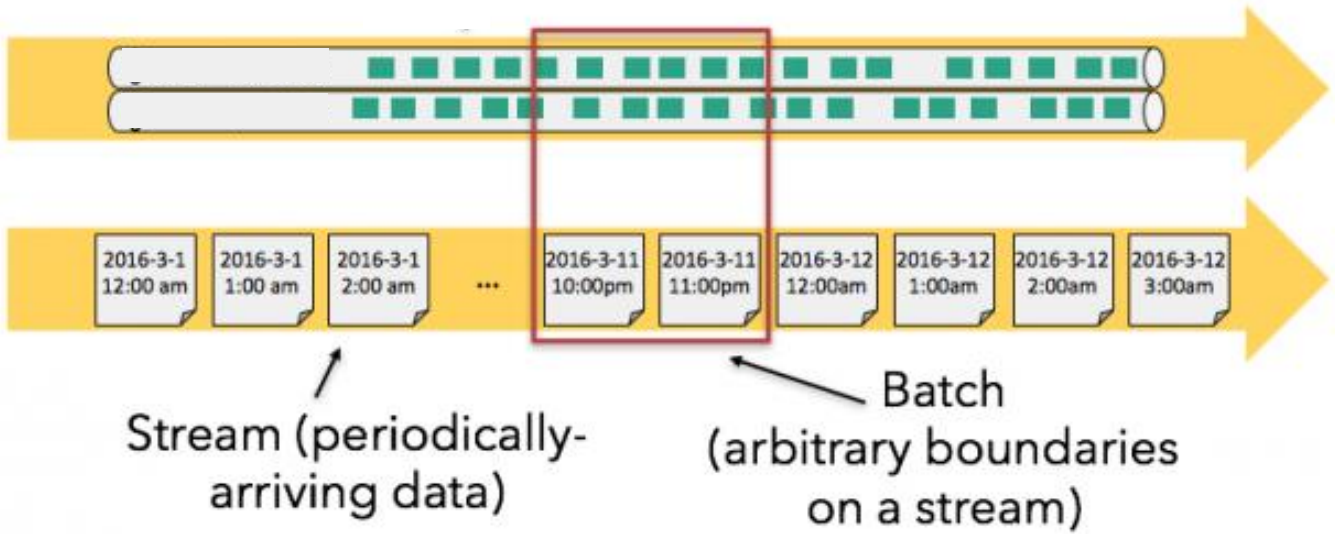
STREAMING TYPES | EVENT BASED

- Один event в промежуток времени
- Параллельные, но независимые
- Задержка ~10ms



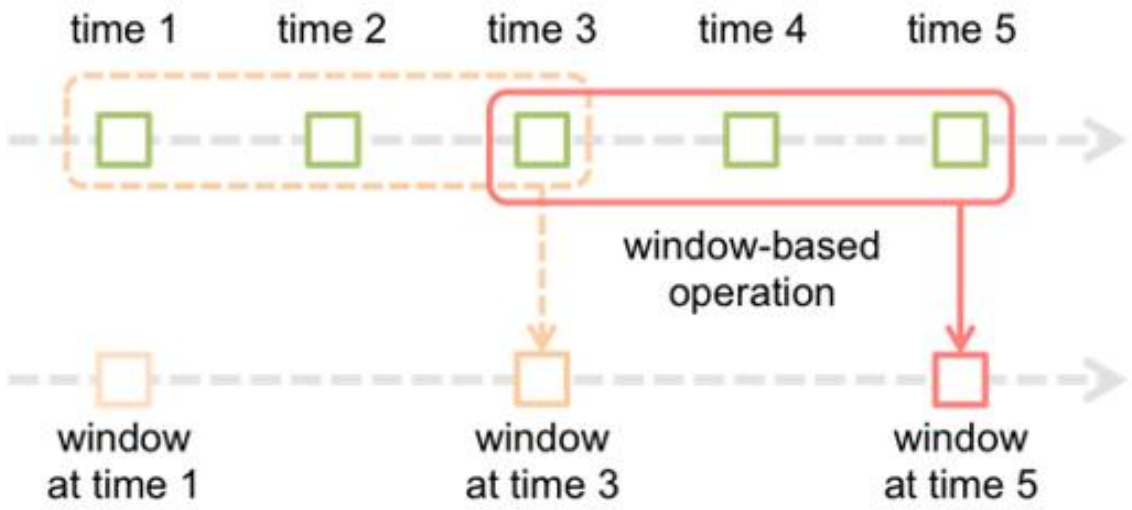
STREAMING TYPES | MICRO-BATCH

- Зависимость от размера батча
- Батчи выполняются последовательно
- Задержка >>1s



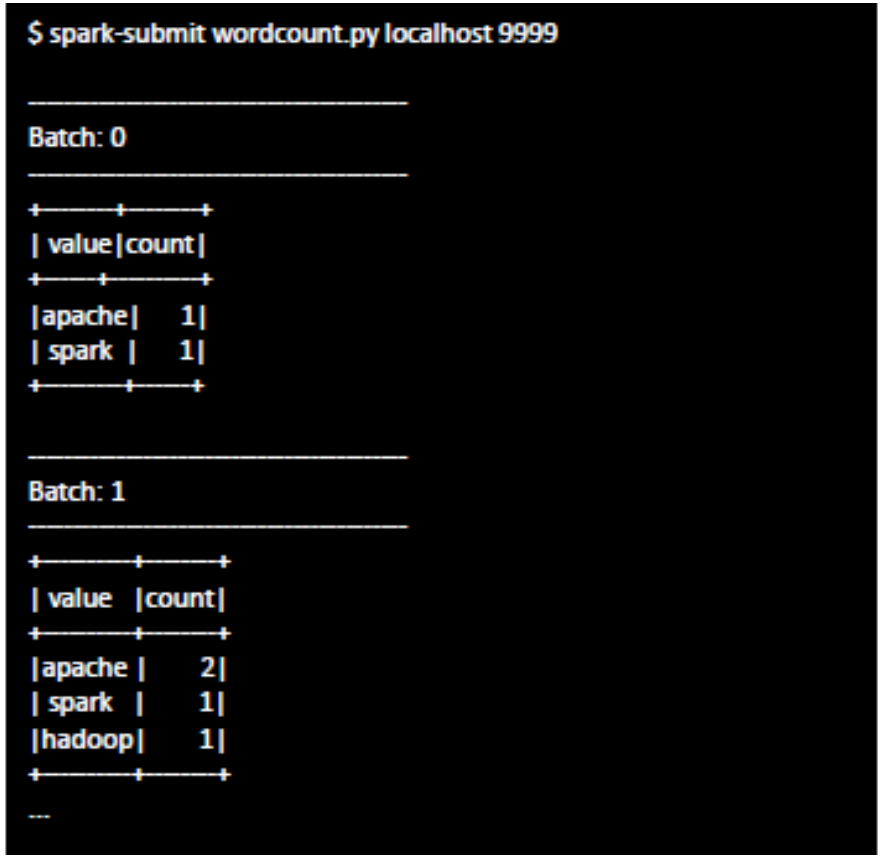
STREAMING TYPES | WINDOWS

- Окно – вид батча

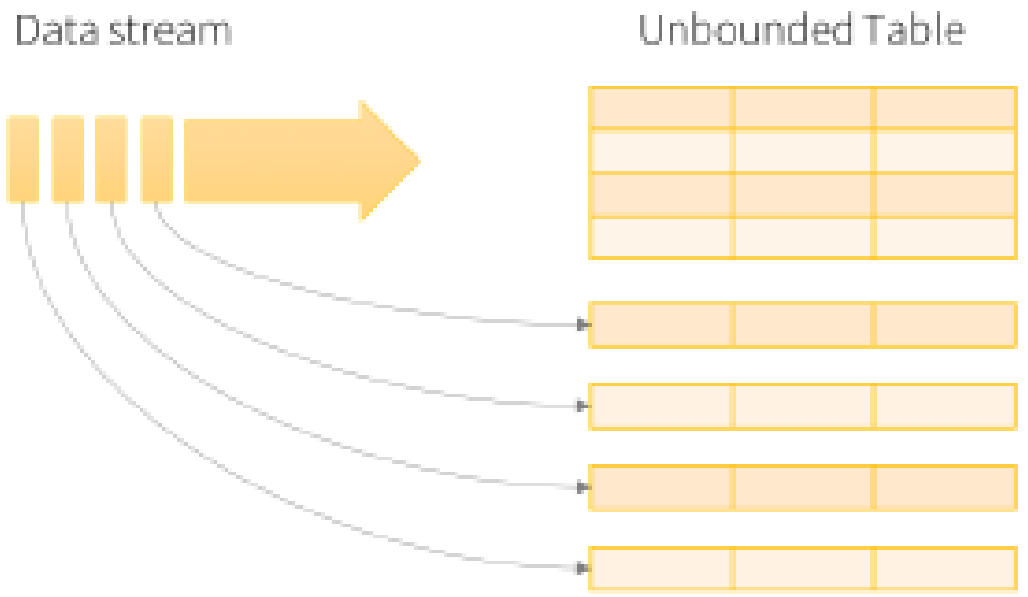


STREAMING – ОСНОНВАЯ ИДЕЯ

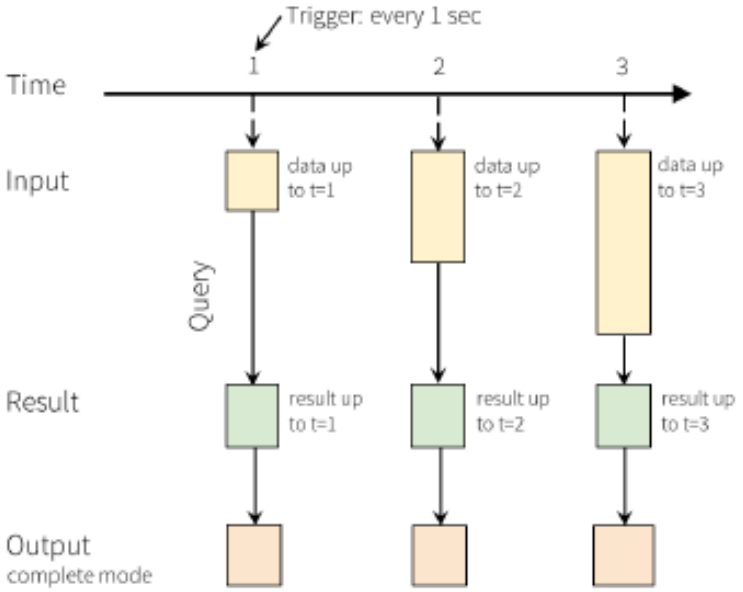
```
from pyspark.sql import SparkSession
from pyspark.sql.functions import explode
from pyspark.sql.functions import split

spark = SparkSession.builder.appName("WordCount").getOrCreate()

lines = spark.readStream.format("socket").option("host",
"localhost").option("port", 9999).load()

words = lines.select(explode(split(lines.value, " ") ).alias("word") )

wordCounts = words.groupBy("word").count()
```



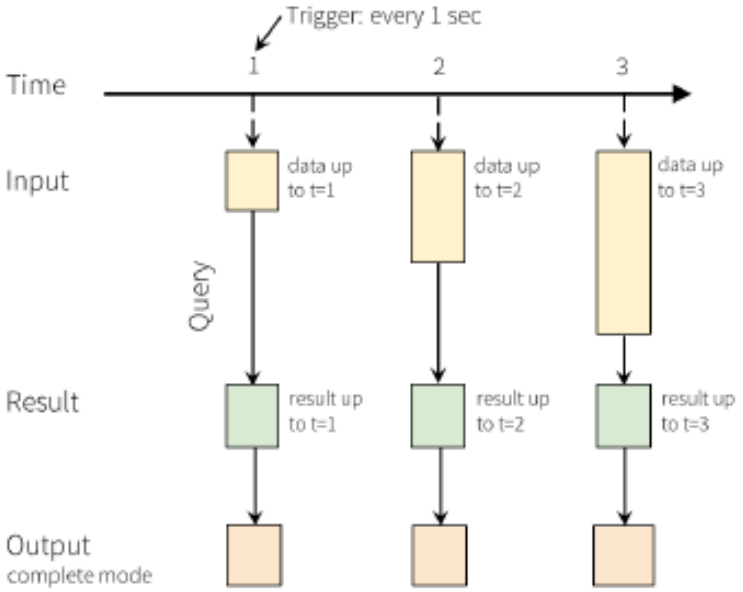
STREAMING – ОСНОНВАЯ ИДЕЯ

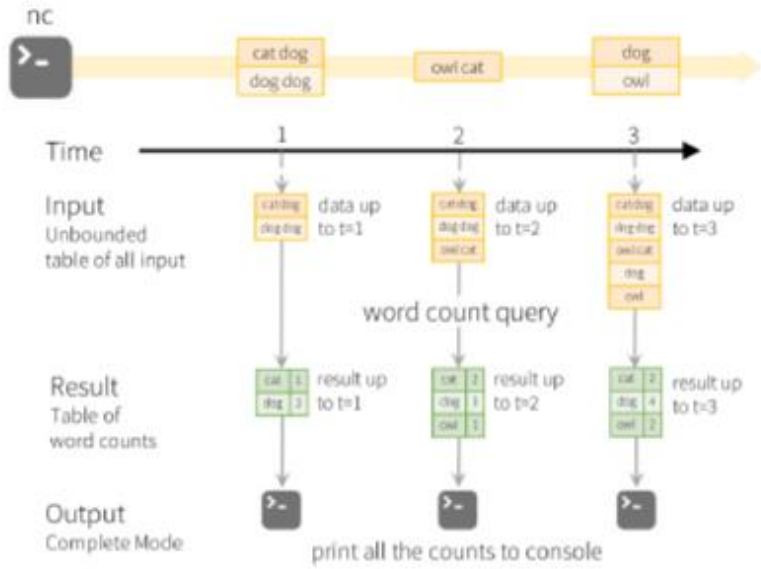


STREAMING – ОСНОНВАЯ ИДЕЯ



STREAMING – ОСНОНВАЯ ИДЕЯ

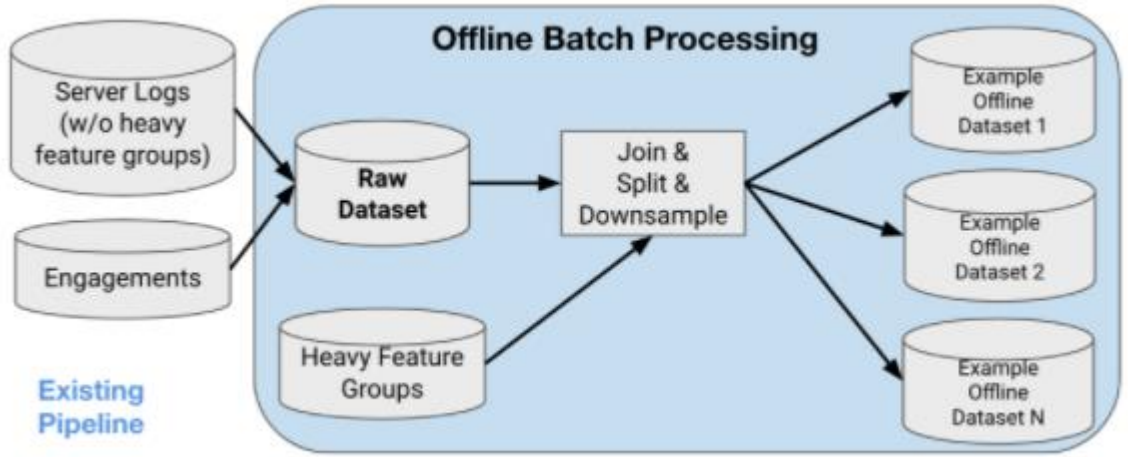




STREAMING CASE

CASE

- Мы можем получить TOP твитов в вашем аккаунте
- Эти твиты выводятся вам с помощью рекомендательной системы (онлайн)
- Сейчас модель переобучается 1 раз в 7 дней, задача переобучать 1 раз в день



CASE

- Почему надо переобучить?

Score падает с каждым днем!

	T-4	T-3	T-2	T-1
RCE (Relative Cross Entropy)	*38.81*	*38.82*	*38.84*	*39.12*

RCE?

CASE

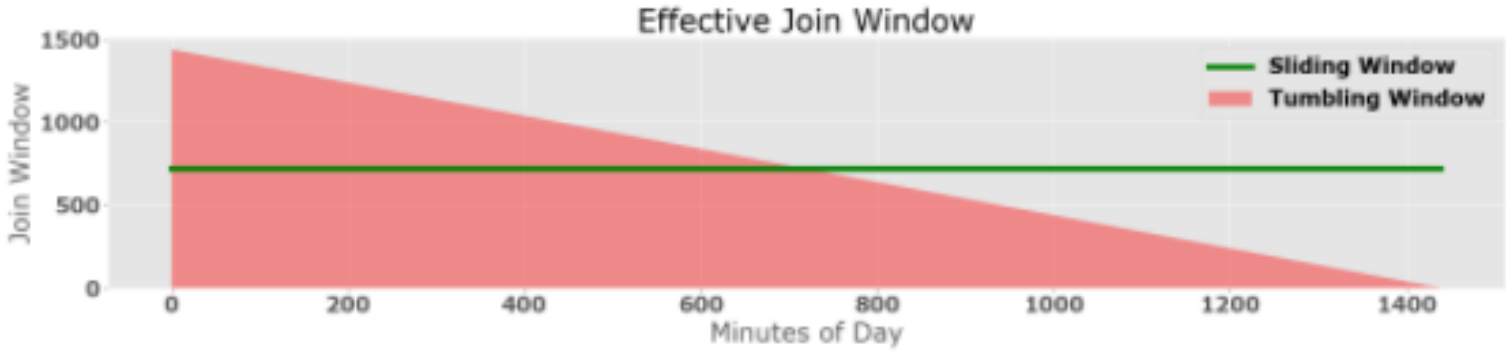
- Почему надо переобучить?

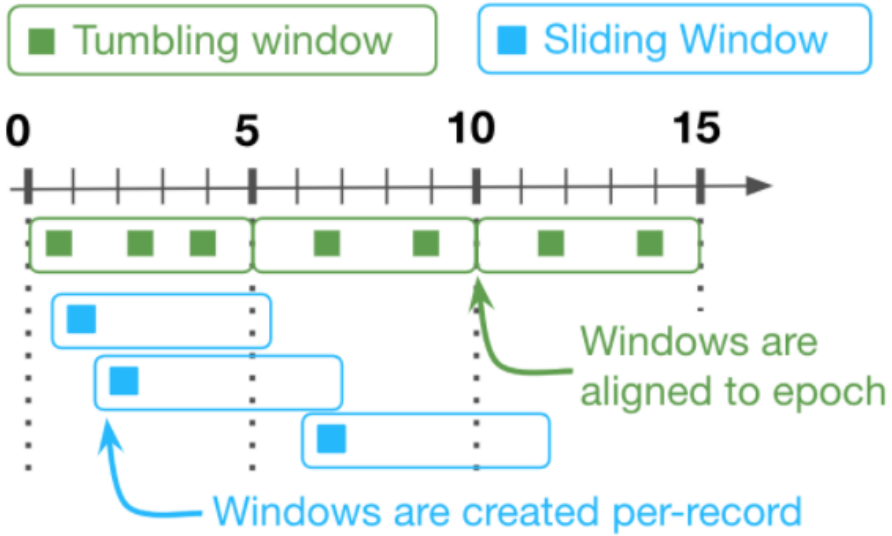
  Score падает с каждым днем!

	T-4	T-3	T-2	T-1
RCE (Relative Cross Entropy)	*38.81*	*38.82*	*38.84*	*39.12*

rce – на сколько предсказание лучше, чем базовая модель «без фич по пользователю» (straw man of naïve prediction)

CASE


Effective Join Window
Sliding Window
Tumbling Window
Join Window
1500
1000
500
0
0
200
400
600
800
1000
1200
1400
Minutes of Day


Tumbling window
Sliding Window
0
5
10
15
Windows are aligned to epoch
Windows are created per-record

ПОПРОБУЕМ САМОСТОЯТЕЛЬНО

SPARK ПРОЕКТ

SPARK ПРОЕКТ

- Задачи реализуемые на Spark ML:

  - Создайте единый набор данных: акции + новости с Reddit

  - Преобразуйте текст в «фичи» на которых сможет обучиться алгоритм

  - Сделайте выбор окна для обучения: 3 дня, 7 дней, 14 дней, 21 день

  - Определите, что вы будете определять: фактическую цену / % изменения