

# APACHE SPARK INTRO FOR DATA SCIENTISTS

О программе



# КОРОТКО О НАС

- Занимаемся обучением работе с данными с 2015 года
- Используем андрагогику и сжатые практикоориентированные программы



# Newprolab

Мы – школа по работе с данными и лаборатория:

- Project Based Education: обучаем работе с инструментами и технологиями на реальных задачах – лабах.
- Предоставляем работу с облачным кластером.
- Соблюдаем баланс между самостоятельностью и поддержкой.

# О ПРОГРАММЕ

# Цели программы

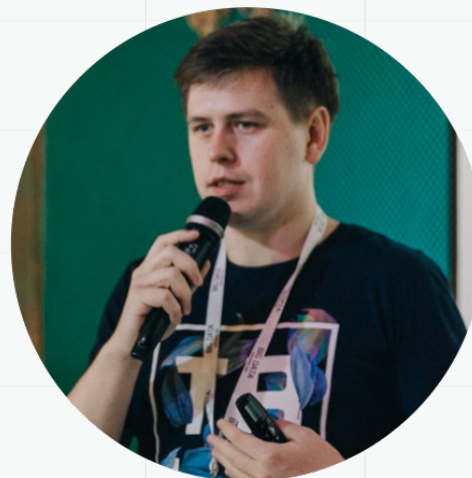
Как перейти к разработке процессов на распределенной системе: распределенные вычисления, основы Spark, Python/Scala для Spark, типичные задачи и кейсы, ML.

# Преподаватели



**Сергей Гришаев**

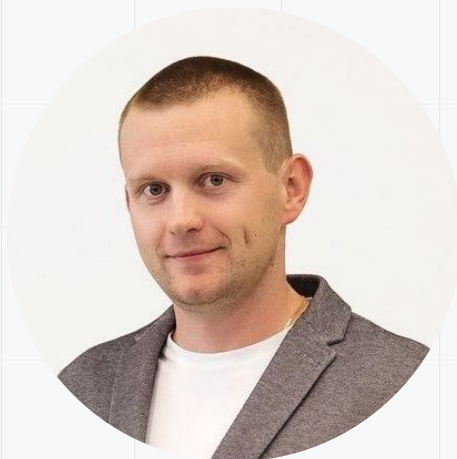
Architect  
**Сбермаркет**



**Егор Матешук**

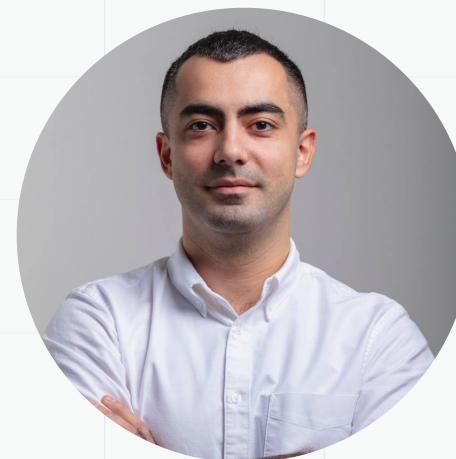
Технический директор  
**ГПМ Дата**

# Координаторы



**Владимир Васев**

Исполнительный директор по аналитике данных  
**ПАО Сбербанк**



**Назим Джавадов**

Senior DS  
**Visa**



# Менеджер программы



**Ольга Сафронова**

Newprolab

# Формат

- 6 недель, 2 занятия в неделю
- Каждое занятие: 3 часа в zoom с 19:00 до 22:00 мск
- 5 лаб и 10 тестов
- Необходимо не менее 10 часов в неделю
- Study Buddy

# Формат

Можно объединиться в команду до 4-х человек и выполнять лабы вместе (чекером проверяем индивидуально) – обсуждать, помогать, поддерживать

Можно делать индивидуально. Но вместе интереснее!



# ИНФРАСТРУКТУРА



# Ресурсы

1. Общий на всех кластер со Spark 2.4.7. Конфигурация: 18 нод по 16 CPU, 80GB RAM. 2 мастера с 32 ядрами и 256GB RAM.
2. Доступ к кластеру по SSH и через JupyterHub.
3. Личный кабинет с календарем занятий и чекерами для лаб.
4. GitHub с материалами занятий.
5. Чат с топиками в Телеграме
6. Записи занятий с тайм-метками в СберУниверситете

# УСПЕШНОСТЬ ПРОХОЖДЕНИЯ



# Удостоверение

## О ПОВЫШЕНИИ КВАЛИФИКАЦИИ

Настоящее удостоверение свидетельствует о том, что

с 3 октября 2022 г. по 12 декабря 2022 г. прошел(а) обучение

в АНО ДПО «Корпоративный университет Сбербанка»

по дополнительной профессиональной программе повышения квалификации

**Apache Spark для задач дата инжиниринга**

в объеме 116 академических часов

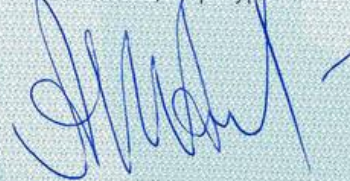
ПК-200671

Регистрационный номер 22-219772

Дата выдачи 12.12.2022

Город Москва

Директор по учебно-методической работе –  
начальник центра уровней программ



Шаталов А.И.







# Certificate of Completion

WITH HONOURS



this is to certify that

**ALEXANDER KRAEV**

has demonstrated exceptional dedication to mastering the science and technology of

**APACHE SPARK FOR DATA ENGINEERING**

a 5-week course that covers Apache Spark, Scala Api, Spark Dataframes,  
Spark Streaming, Spark UI

24.11.2022

Date of Issue

A handwritten signature in blue ink, likely belonging to Anna Zhovtis.

Anna Zhovtis  
CEO, New Professions Lab

Certificate number: 30792815881535266206177170405



# Успешность

**Для успешного прохождения** необходимо правильно ответить суммарно минимум на 60% вопросов тестов и сделать любые 3 лабы из 5.

**С отличием** – правильно ответить суммарно минимум на 60% вопросов тестов и сделать 5 лаб из 5.



# ТЕМЫ И ЗАНЯТИЯ

## Занятия

№	Дата	Лекция	Спикер
1	8 июня, четверг, 19:00 мск	Git, pycharm и intelliJ	Сергей Гришаев
2	13 июня, вторник, 19:00 мск	Введение в распределенные вычисления	Сергей Гришаев
3	15 июня, четверг, 19:00 мск	Spark RDD	Сергей Гришаев
4	20 июня, вторник, 19:00 мск	Spark Dataframes I	Сергей Гришаев
5	22 июня, четверг, 19:00 мск	Hadoop для Spark-пользователя	Егор Матешук
6	27 июня, вторник, 19:00 мск	Python/Scala для Spark	Сергей Гришаев
7	29 июня, четверг, 19:00 мск	Spark Dataframes II	Сергей Гришаев
8	4 июля, вторник, 19:00 мск	Введение в Spark Streaming	Сергей Гришаев
9	6 июля, четверг, 19:00 мск	Принципы распределенного ML	Сергей Гришаев
10	11 июля, вторник, 19:00 мск	Spark ML: основы и создание пайплайнов	Сергей Гришаев
11	13 июля, четверг, 19:00 мск	Практический ML на Spark	Сергей Гришаев
12	18 июля, вторник, 19:00 мск	Разбор лаб и Q&A	Сергей Гришаев

# Принципы

- Занятия: задавайте вопросы (не существует глупых вопросов).
- Лабы: просите помощи у сокурсников и координаторов, но вначале попробуйте решить сами.
- Делайте лабы заранее. В последний момент можно не успеть.
- **Делайте лабы самостоятельно.**



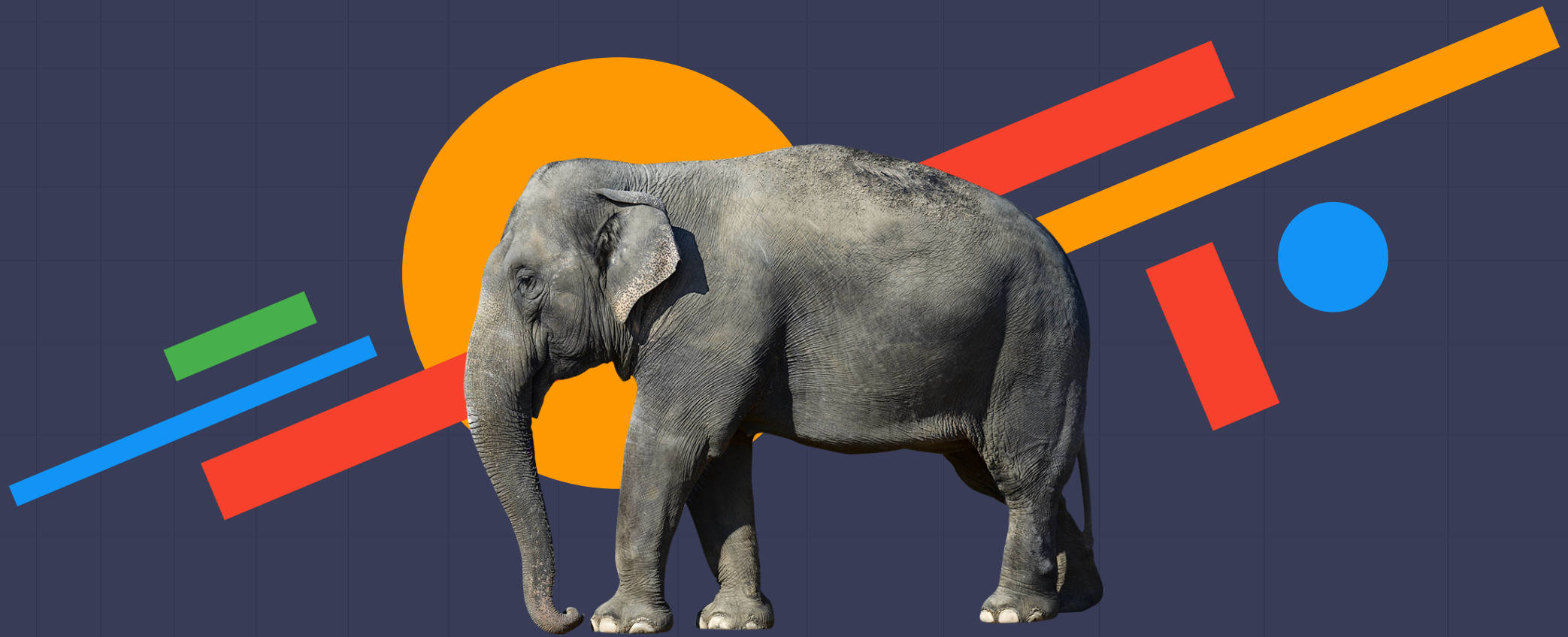
# ЛАБЫ

## Лабораторные

№	Лаба	Лаба открывается	Дедлайн чекера
1	Лаба 1. Дескриптивный анализ рейтингов фильмов на RDD	15 июня	27 июня
2	Лаба 2. Content-based рекомендательная система на Dataframes	27 июня	5 июля
3	Лаба 3. Создание витрины данных из разных источников: файлы, NoSQL-хранилища, реляционные базы данных	29 июня	13 июля
4	Лаба 4. Дообучение модели на новых данных по расписанию, инференс модели в real-time	10 июля	18 июля
5	Лаба 5. Соревнование: рекомендация фильмов на основе данных об истории телесмотрения	13 июля	18 июля



# ВАШИ ВОПРОСЫ



# Big Data is Love

NEWPROLAB.COM