Занятие 4

Изучение знаний и опыта, полученных в результате проекта ЕФО «Большие данные для аналитики рынка труда»

Обзор технического построения системы сбора и анализа данных онлайн-вакансий: от изучения ландшафта источников данных до их визуализации

Мауро Пелуччи

Ноябрь 2021

Темы

1. Обзор и краткое повторение
2. Что такое конвейер?
3. Уровень хранения
4. Фреймворк Spark
5. Лабораторная работа
    1. Поиск новых наименований должностей
    2. Поиск новых профессий

Темы

1. **Обзор и краткое повторение**
2. Что такое конвейер?
3. Уровень хранения
4. Фреймворк Spark
5. Лабораторная работа
    1. Поиск новых наименований должностей
    2. Поиск новых профессий

Общий поток данных

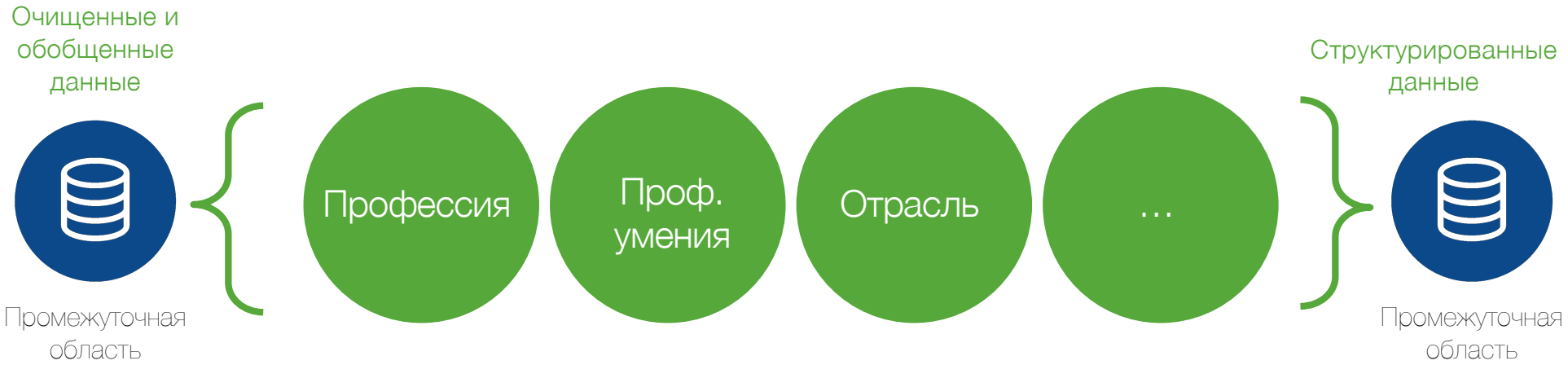
Извлечение информации

- Цель:
  - Извлечь и структурировать информацию из данных для передачи на уровень представления
- Задачи:
  - Обработка огромного массива неоднородных данных на разных языках
- Подход:
  - Разработать адаптируемую схему, подстроенную к различным особенностям информации. Некоторые актуальные задачи:
    - Классификация по **профессии**: комбинированные методы, такие как машинное обучение, тематическое моделирование, обучение без учителя
    - Классификация по **профессиональным умениям**: другие различные комбинированные методы, такие как анализ текста с учетом сходства на основе корпуса или знаний
- Особенности:
  - Гарантировать извлечение объяснимой информации, регистрацию методов классификации и соответствующие особенности.

Извлечение и классификация информации Аналитика рынка труда в реальном времени

Извлечение информации — это сфера обработки естественного языка, которая связана с поиском фактической информации в произвольном тексте.

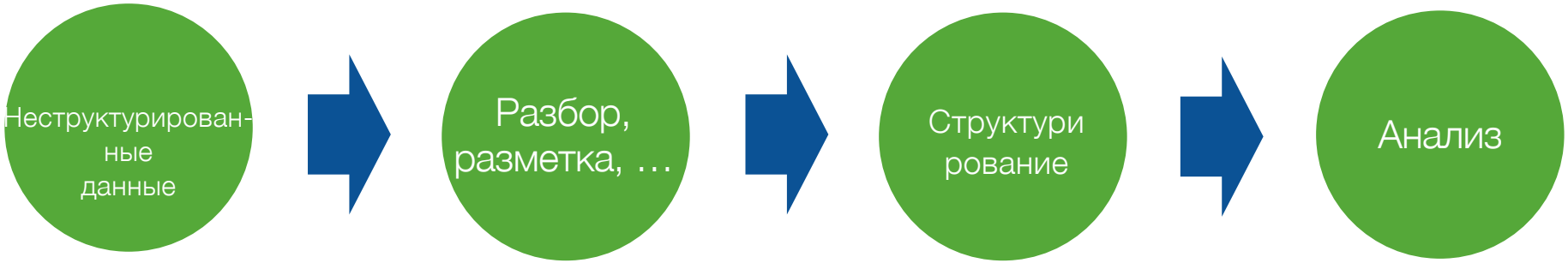
Для выполнения этой задачи используются методы машинного обучения (обучение на основе онтологии, обучение с учителем и без учителя), чтобы сопоставлять объявления о работе со стандартными классификациями.



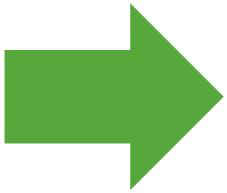
Машинное обучение → Обучение на основе онтологий, обучение с учителем и без учителя, и др.

Извлечение информации

Извлечение информации: анализ неструктурированного документа с целью извлечения определенной информации.



Вакансия



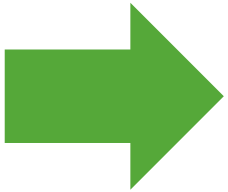
Извлечение информации

- Профессия
- Профессиональные умения
- Время
- Регион
- Отрасль
- …

Младший разработчик ПО

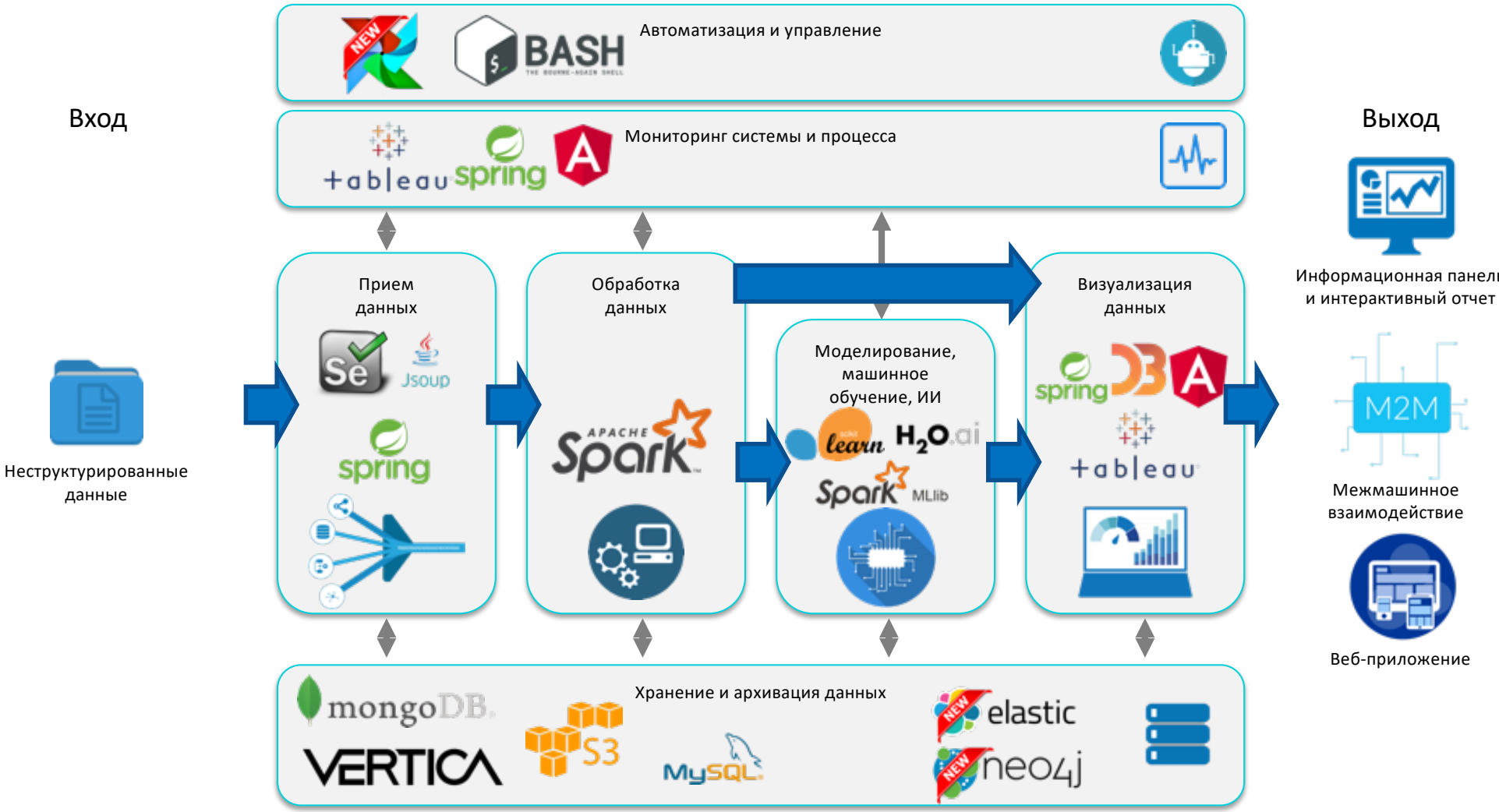
В качестве младшего разработчика ПО, вы будете разрабатывать прекрасное программное обеспечение для использования в сфере сопоставления полей, сбора данных, сенсорных сетей, уличной навигации, и многих других. Вы будете сотрудничать с другими программистами и разработчиками в целях самостоятельного проектирования и внедрения высококачественных веб-приложений, API, соответствующих ограничениям REST, и обеспечения интеграции сторонних решений.

Мы ищем увлеченного, преданного делу разработчика, умеющего решать и формулировать сложные задачи в сфере проектирования приложений, разработки и взаимодействия с пользователем. Работа в нашем офисе в Харвелле, Великобритания.



Извлечение информации

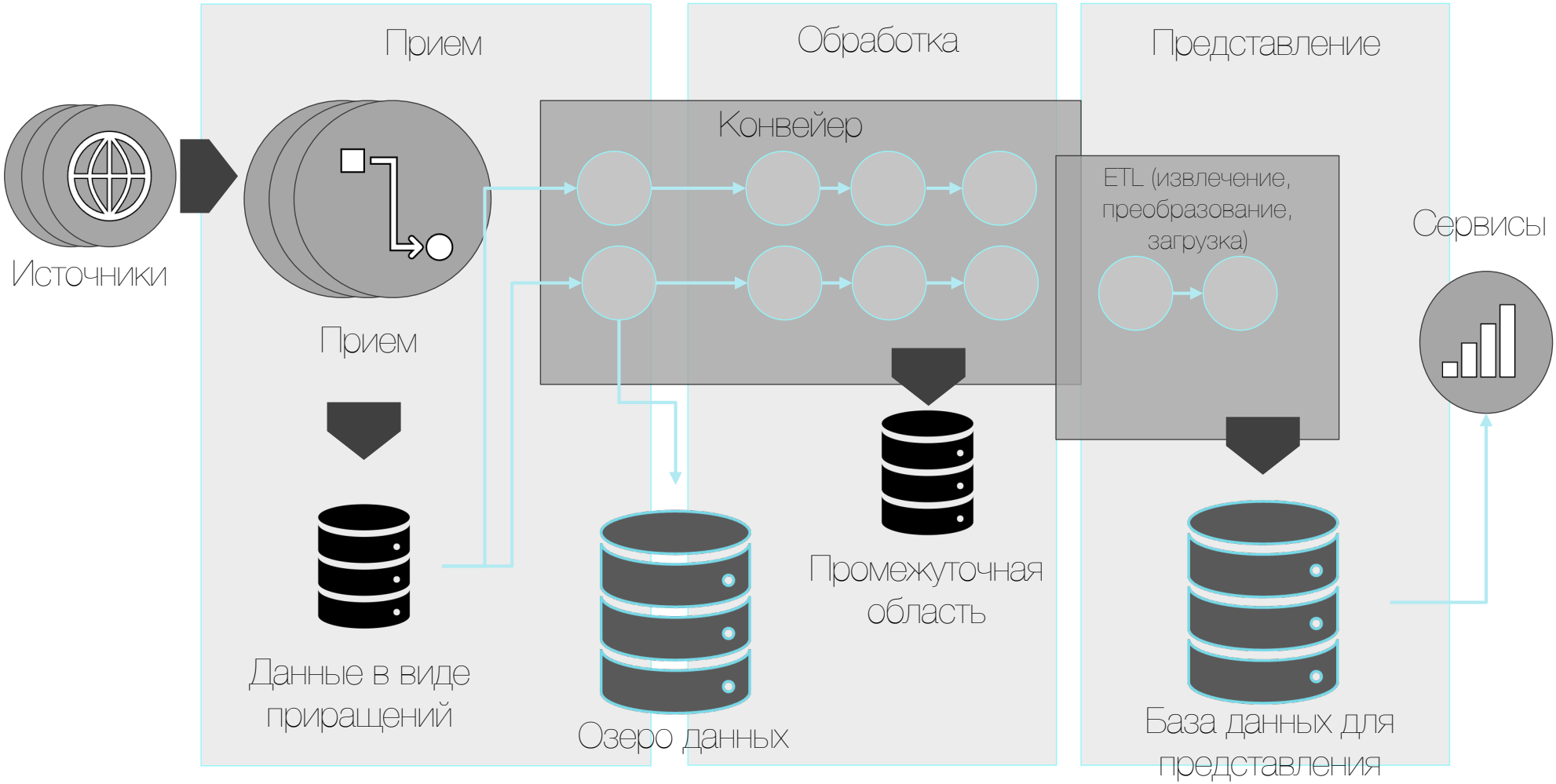
- 2512 — Разработчик ПО
- Проф. умения: разработка ПО, внедрение веб-приложений, решение проблем, разработка пользовательских взаимодействий
- Харвелл, Великобритания
- …


Вход
Неструктурированные данные
Автоматизация и управление
Мониторинг системы и процесса
Прием данных
Обработка данных
Моделирование, машинное обучение, ИИ
Визуализация данных
Хранение и архивация данных
Выход
Информационная панель и интерактивный отчет
Межмашинное взаимодействие
Веб-приложение

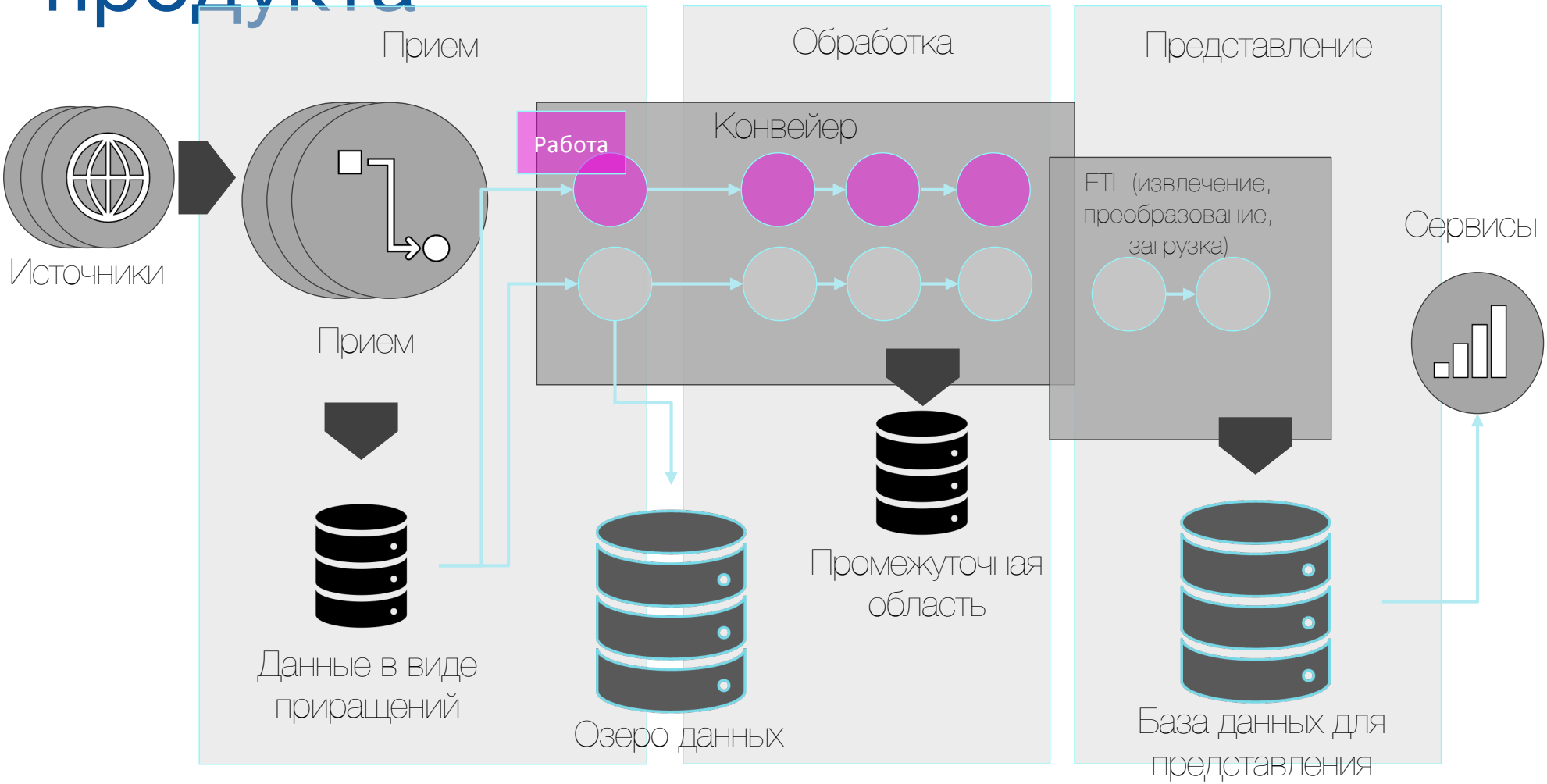
Темы

1. Обзор и краткое повторение
2. **Что такое конвейер?**
3. Уровень хранения
4. Фреймворк Spark
5. Лабораторная работа
    1. Поиск новых наименований должностей
    2. Поиск новых профессий

Анатомия информационного продукта

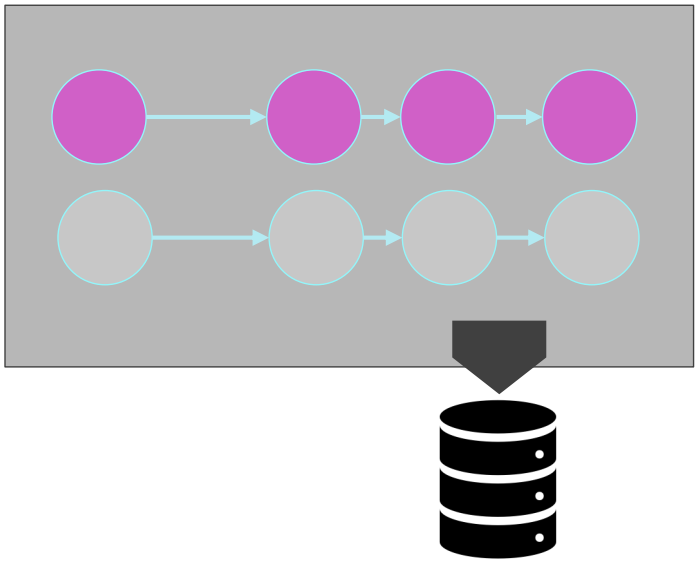


Анатомия информационного продукта



Конвейер обработки данных

Еще одна компьютерная программа



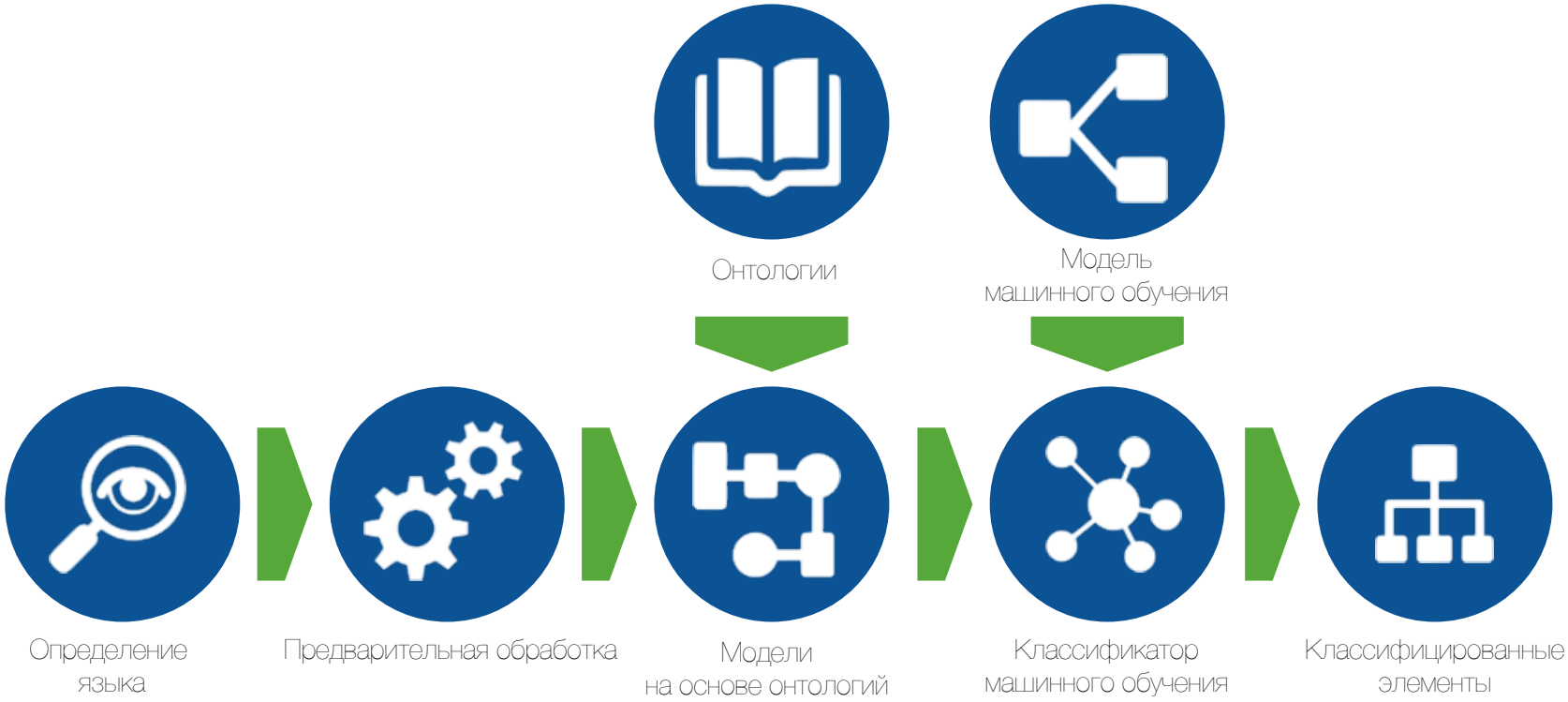
Промежуточная область

Пакетное задание

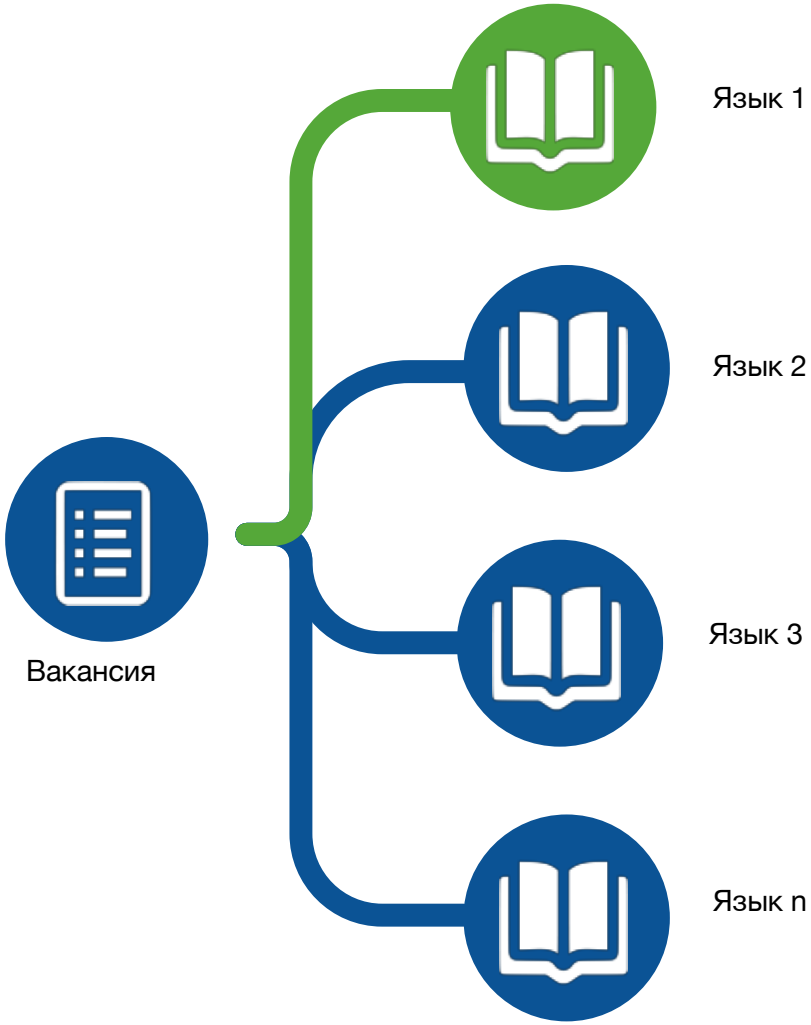
Работа == функция ([входящий набор данных]):
[выходящий набор данных]

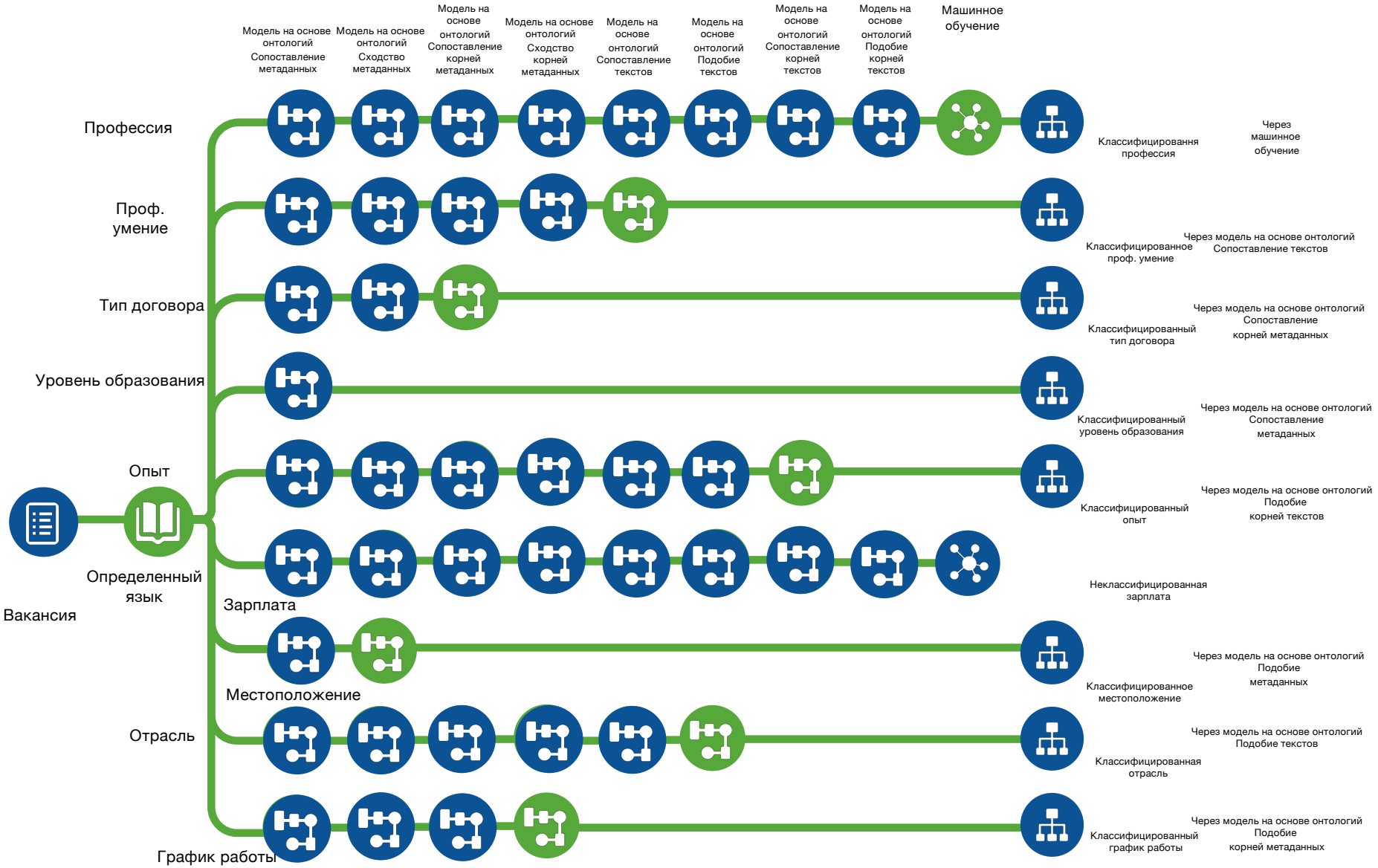
- Поддается проверке
- Элементарное
- Детерминированное
- Идемпотентное
- Отсутствие других вводных параметров

Схема конвейера



Конвейер и определение языка

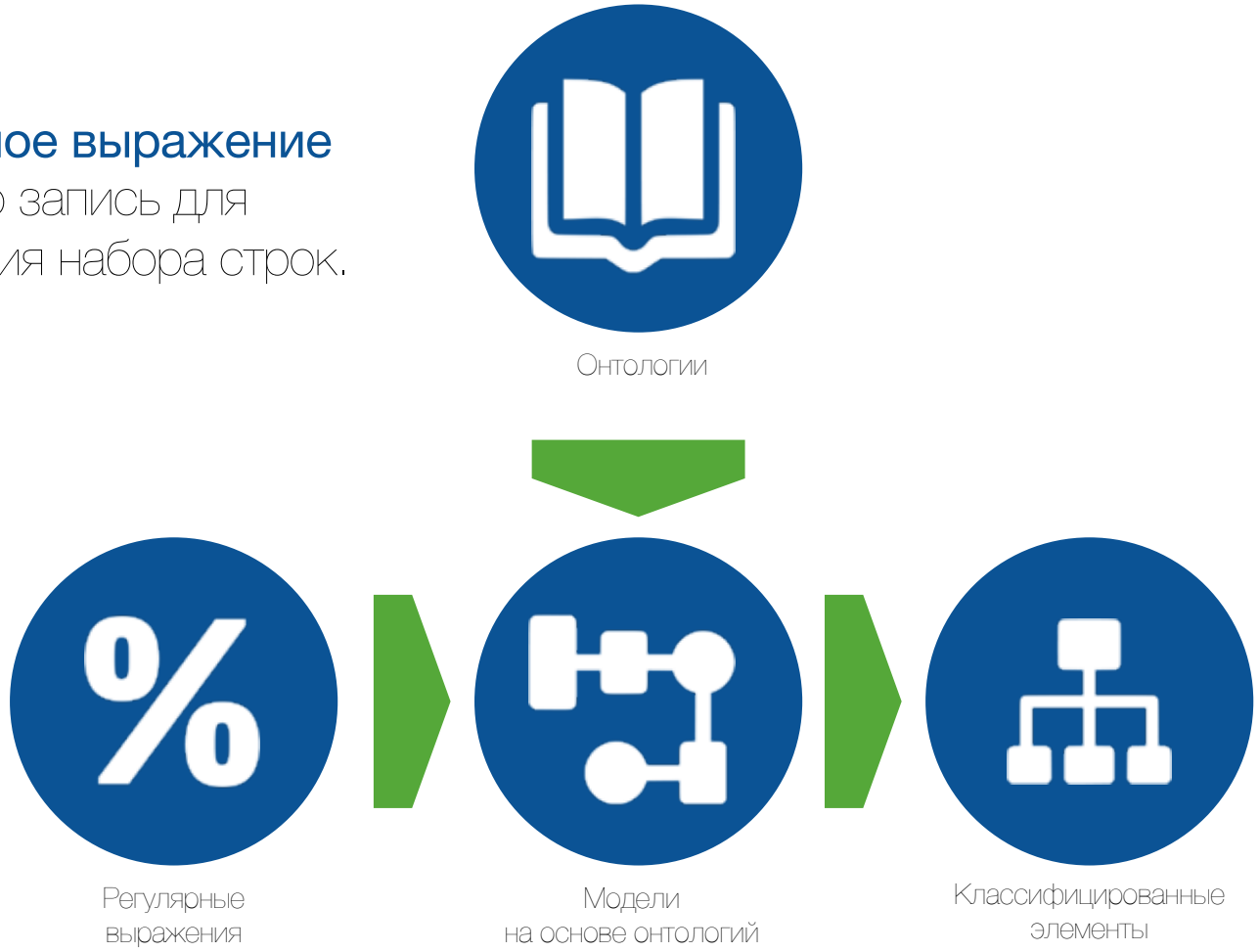



Модель на основе онтологий Сопоставление метаданных
Модель на основе онтологий Сходство метаданных
Модель на основе онтологий Сопоставление корней метаданных
Модель на основе онтологий Сходство корней метаданных
Модель на основе онтологий Сопоставление текстов
Модель на основе онтологий Подобие текстов
Модель на основе онтологий Сопоставление корней текстов
Модель на основе онтологий Подобие корней текстов
Машинное обучение
Профессия
Проф. умение
Тип договора
Уровень образования
Опыт
Определенный язык
Вакансия
Зарплата
Местоположение
Отрасль
График работы
Классифицированная профессия
Через машинное обучение
Классифицированное проф. умение
Через модель на основе онтологий Сопоставление текстов
Классифицированный тип договора
Через модель на основе онтологий Сопоставление корней метаданных
Классифицированный уровень образования
Через модель на основе онтологий Сопоставление метаданных
Классифицированный опыт
Через модель на основе онтологий Подобие корней текстов
Неклассифицированная зарплата
Классифицированное местоположение
Через модель на основе онтологий Подобие метаданных
Классифицированная отрасль
Через модель на основе онтологий Подобие текстов
Классифицированный график работы
Через модель на основе онтологий Подобие корней метаданных

Компоненты на основе онтологий

Регулярные выражения

Регулярное выражение — это запись для определения набора строк.



Регулярное выражение для определения зарплаты

Тестирование (одна работа)

Как тестировать конвейер обработки данных?

- Тестировать одну работу / один компонент
- Стандартный набор данных («золотой» набор данных или имитация набора данных)
  - Сгенерировать входные данные
  - Запустить в рамках локального / маленького кластера
  - Проверить выходные данные

Что нам нужно?
Набор средств

Статистические методы

Инструменты

Исследование механизма взаимодействия с пользователем

Анализ временных рядов

Подстановка расчетных данных вместо недостающих

Интеллектуальный анализ данных

Многоуровневое моделирование

Классификация и создание кластеров

Распознавание закономерностей

A/B-тестирование

Главный компонент и факторный анализ

Машинное обучение

Прогнозирование

Анализ сети

Методы регрессии

Что нам нужно?
Набор средств

Статистические методы | **Инструменты** | Исследование механизма взаимодействия с пользователем

Языки
Python
R
Scala
SQL

Библиотеки
Pandas
Sklearn
OpenNLP
Spacy
Fasttext
Word2Vec
H20.ai
…

Инженерия данных
Hadoop
Spark
Профилирование
ETL (извлечение, преобразование, загрузка)
Объявления о работе
API
Конвейер обработки оптимизированных данных
Хранение оптимизированных данных/ доступ к ним

Облако (AWS)
CI/CD

Визуализация
D3.js
Gephi
R
Matplot
Shiny
Tableau

Что нам нужно?
Набор средств

Статистические методы

Инструменты

Исследование механизма взаимодействия с пользователем

Интерактивное прототипирование

Создание карты сервисного сценария

Наблюдение за пользователем

Создание карты пути пользователя

Темы

1. Обзор и краткое повторение
2. Что такое конвейер?
3. **Уровень хранения**
4. Фреймворк Spark
5. Лабораторная работа
    1. Поиск новых наименований должностей
    2. Поиск новых профессий

Ключевые понятия

- Столбчатые форматы данных
- Delta lake

Понятия
Столбчатые форматы данных

- Фильтры — это не единственные «предикаты», которые могут быть спущены
- Для этой цели также может быть использован выбор столбцов
  - В такой базе данных, как PostgreSQL, это можно сделать с помощью оператора SELECT
  - Для файлов нам требуется столбчатый формат файлов
- Данные хранятся по столбцам, а не по строкам
    - Parquet и ORC
    - Формат Delta lake: Delta.io, Hudi, Iceberg
- В сравнении с построчными форматами файлов, в которых данные хранятся по строкам
  - CSV, TSV, JSON, и AVRO

Понятия
Пример: Столбчатый формат в сравнении с построчным

Построчный

	имя	цвет	город	возраст
Строка 1	Том	красный	Чикаго	32
Строка 2	Салли	синий	Париж	87
Строка 3	Майк	зеленый	Лондон	20
Строка 4	Мэри	желтый	Фресно	55

Читает строку № 1

Столбчатый

	Строка 1	Строка 2	Строка 3	Строка 4
имя	Том	Салли	Майк	Мэри
цвет	красный	синий	зеленый	желтый
город	Чикаго	Париж	Лондон	Фресно
возраст	32	87	20	55

Читает столбец «имя»

Что такое **Delta Lake**?

Технология, разработанная для использования с Apache Spark с целью создания озер надежных данных

Проект с открытым исходным кодом delta.io

Databricks [документация Delta Lake]

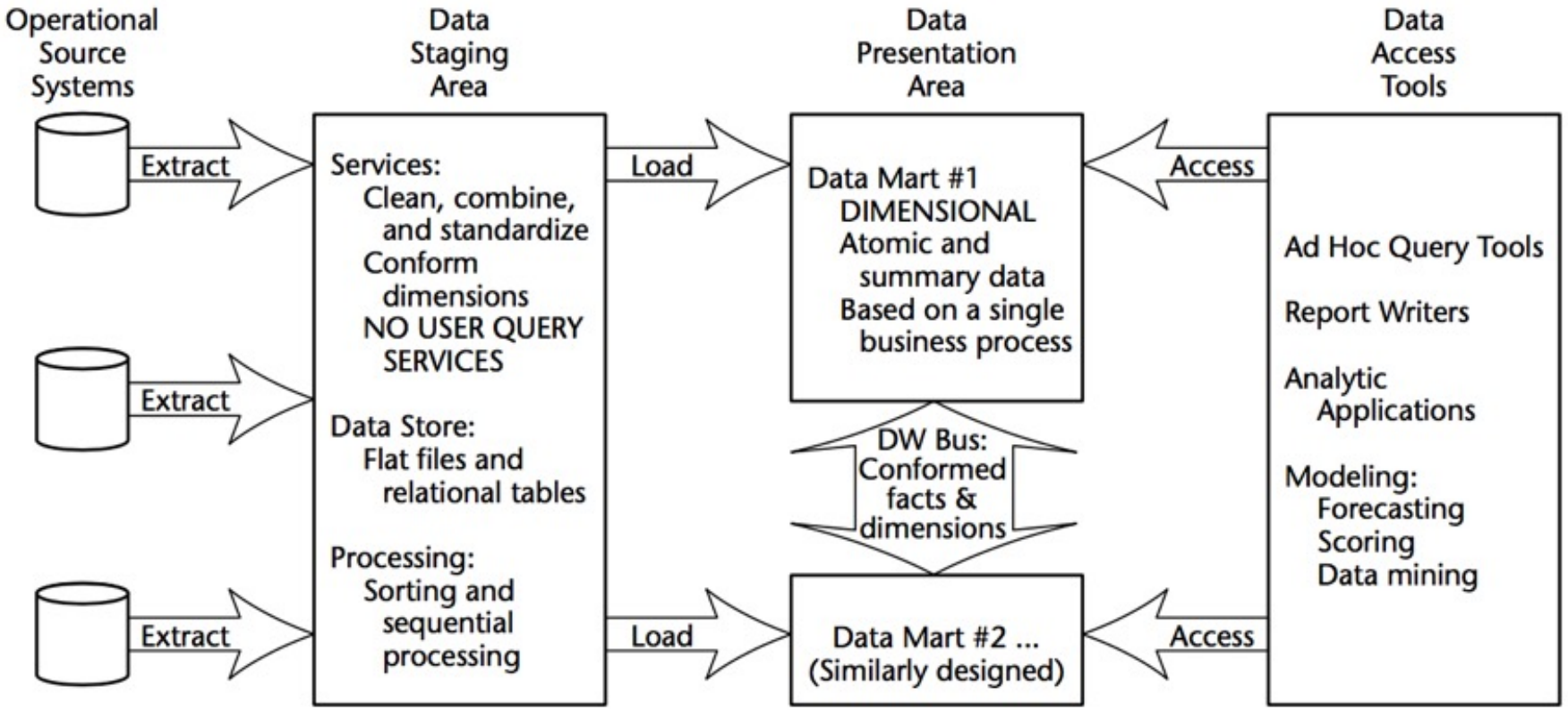
Особенности Delta lake

- Транзакции, соответствующие требованиям ACID, на базе Spark
- Масштабируемая обработка метаданных
- Потоковая передача и унификация пакетов
- Принудительное применение схемы
- Механизм путешествия во времени
- Обновление или вставка и удаление
- Полностью конфигурируемый/ оптимизируемый
- Поддержка структурированной потоковой передачи

Промежуточная область

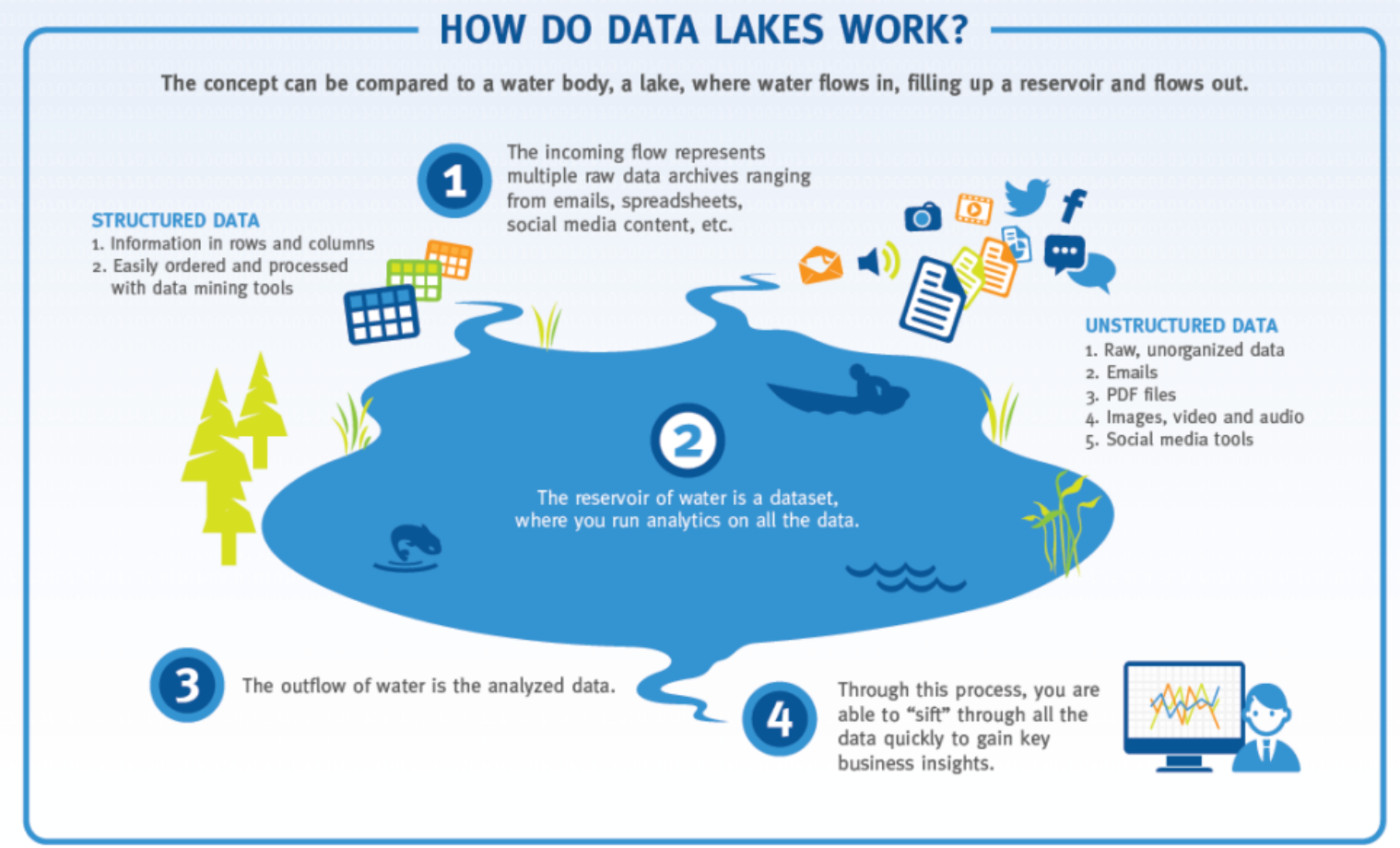
ПРОМЕЖУТОЧНАЯ ОБЛАСТЬ = конвейеры обработки данных, данные и процессы ETL, это как кухня ресторана

- У конечного пользователя не должно быть доступа к данным в промежуточной области: они не готовы к потреблению.
- В промежуточной области проводятся «опасные» операции: очистка данных, поиск и соединение данных, создание витрин данных, …
- Корпоративные пользователи не думают (и не должны думать) о том, что происходит в рамках конвейера обработки данных и процессов ETL.



Кимбалл/Росс

Озеро данных



Источник иллюстрации: EMC
https://40uu5c99f3a2ja7s7miveqgqu-wpengine.netdna-ssl.com/wp-content/uploads/2017/02/Understanding-data-lakes-EMC.pdf

Парадигма озера данных

Хранилище данных

- Агрегированные подмножества
- Представление данных по запросу
- Проверено экспертами
- Структурированное — таблицы, представление данных, отчеты. Узкий контекст
- Качество данных известно и постоянно отслеживается

Озеро данных

- Хранение данных в исходном виде
- Пусть бизнес сам решает, что ему нужно
- Поддержка быстрых изменений
- Возможность отслеживания и визуализации линии данных и истории использования данных
- Неструктурированное — поиск по ключевым словам
- Данные доступны в различных видах: от «сырых» данных до полностью соответствующих требованиям
- Параметры качества зачастую отсутствуют

Современная архитектура озера данных

- Структурирование данных при чтении
- Дескриптивное моделирование данных
- Новые данные могут начать поступать в любое время и будут появляться задним числом
- Гибкость
- Масштабируемость
- Быстрый прием данных
- Подходит для исследований и подхода «снизу вверх»

S3 Bucket
Datalake

Резюме и ключевые слова

- Конвейеры обработки данных и вакансии
  - Еще одна компьютерная программа
  - Пакетное задание
- Разные типы компонентов
  - На основе машинного обучения, на основе онтологий, регулярных выражений, …
- Тестирование конвейера
- Хранилище
  - Разные форматы: json, parquet и delta.io
  - Разные области: метаданные, озеро данных, область представления

Вопросы?

Темы

1. Обзор и краткое повторение
2. Что такое конвейер?
3. Уровень хранения
4. **Фреймворк Spark**
5. Лабораторная работа
    1. Поиск новых наименований должностей
    2. Поиск новых профессий

Де-факто стандартная унифицированная аналитическая платформа для обработки больших данных

Самый крупный проект с открытым исходным кодом в сфере обработки данных

Ключевые понятия и термины

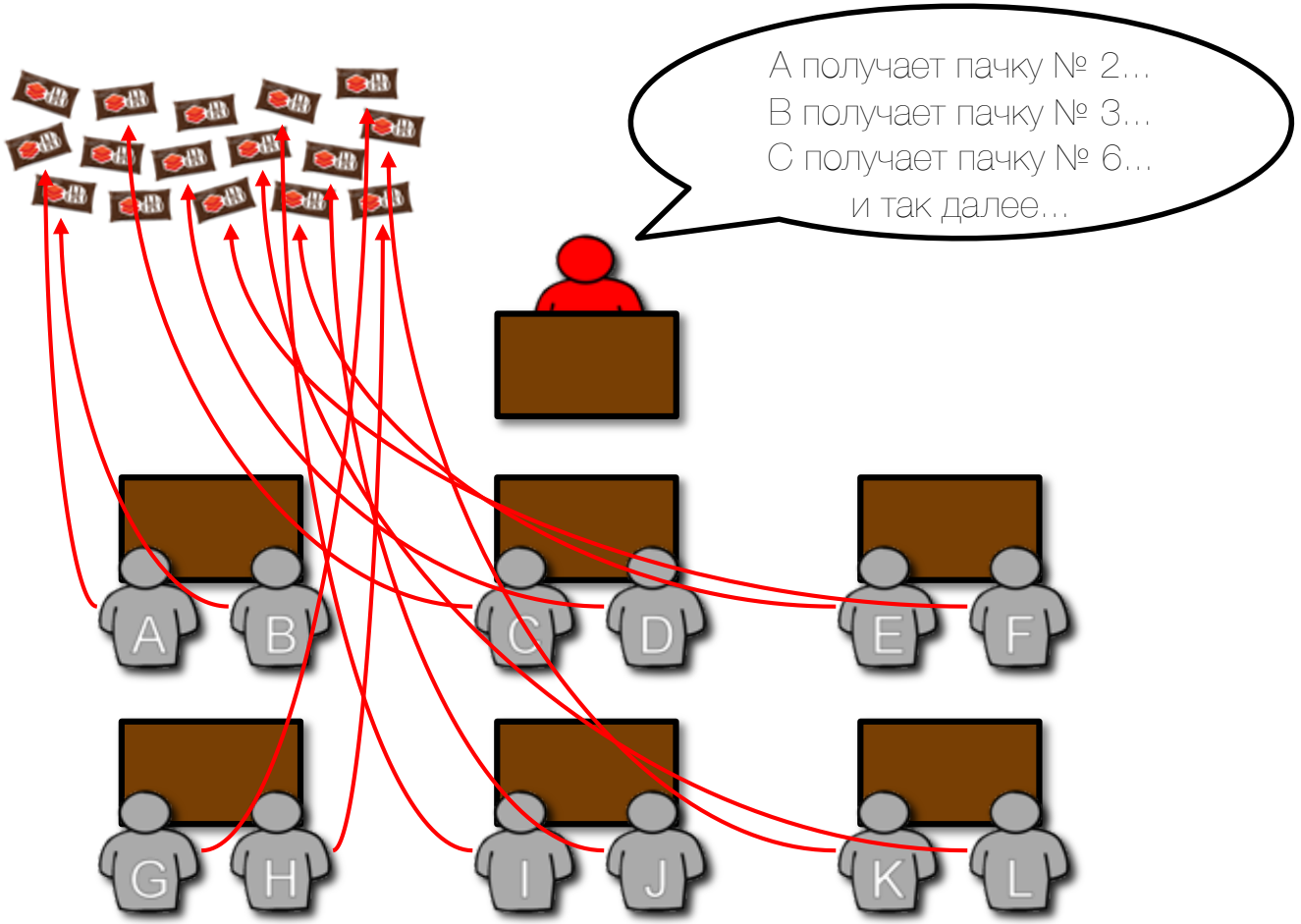
- Общие ресурсы
- Распараллеливание
- Разделы
- Работы, этапы и задачи
- Драйверы
- Исполнители
- Кластер и узлы
- Ядра/потоки выполнения

Сможешь открыть пачку...

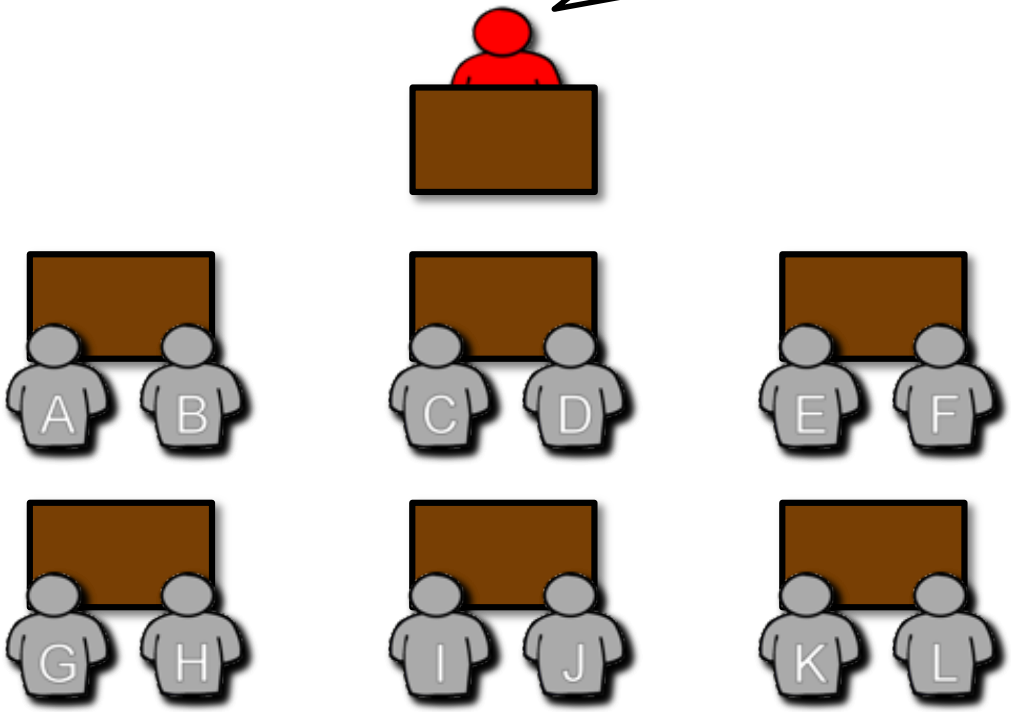
...и съесть все коричневые конфет

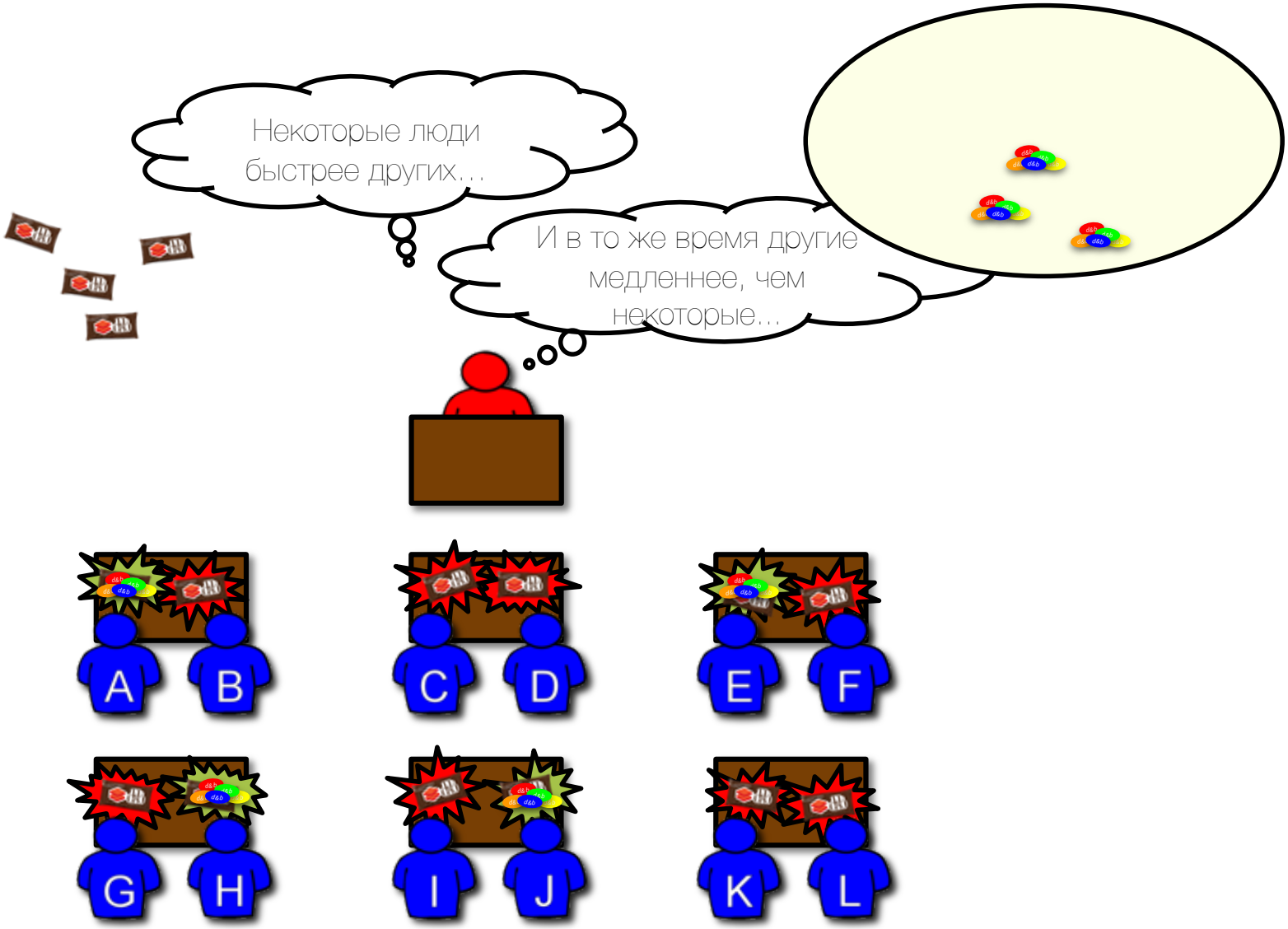
...за 60 секунд?

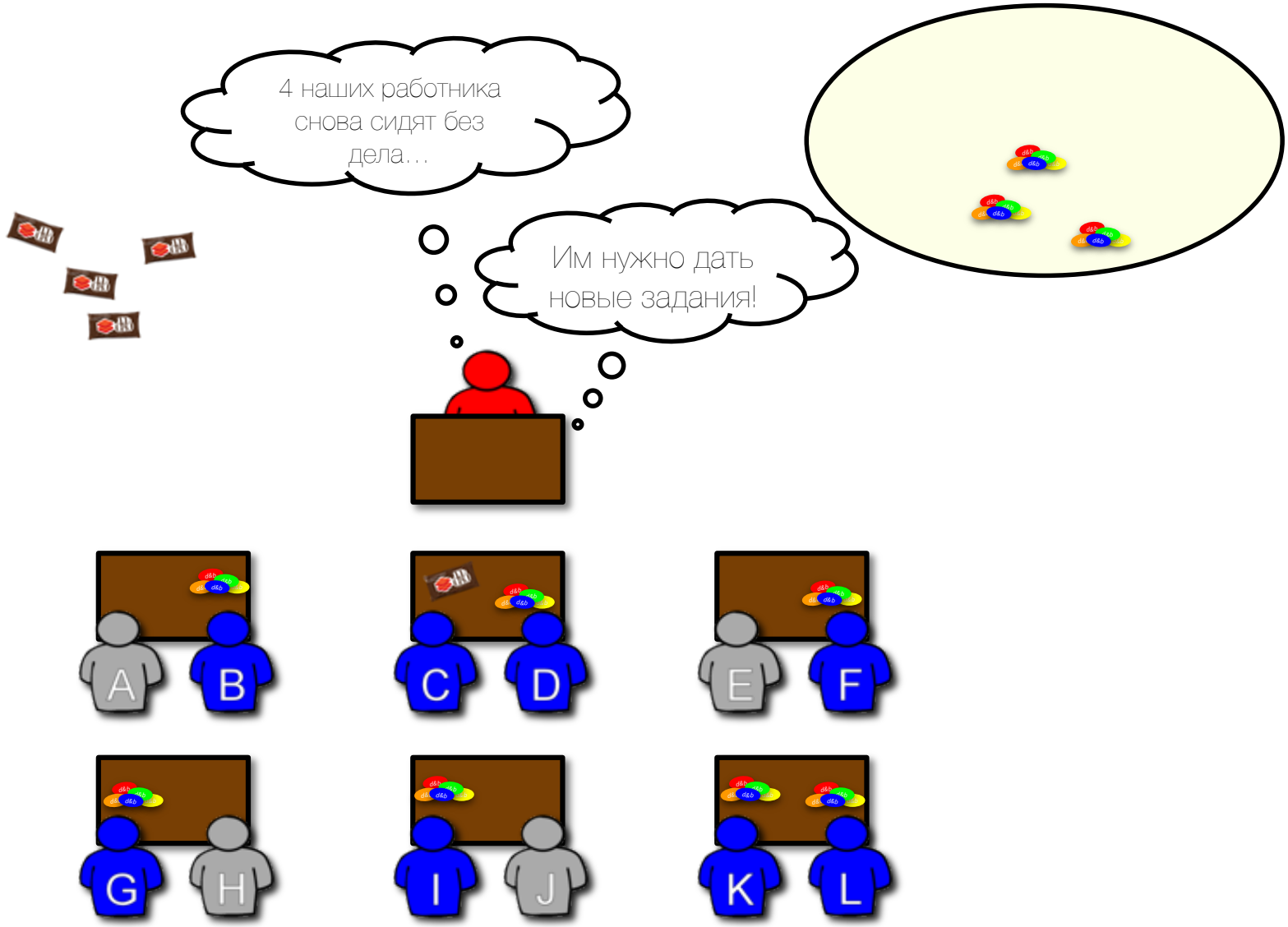
А как насчет 100 пачек M&Ms
за 60 секунд?

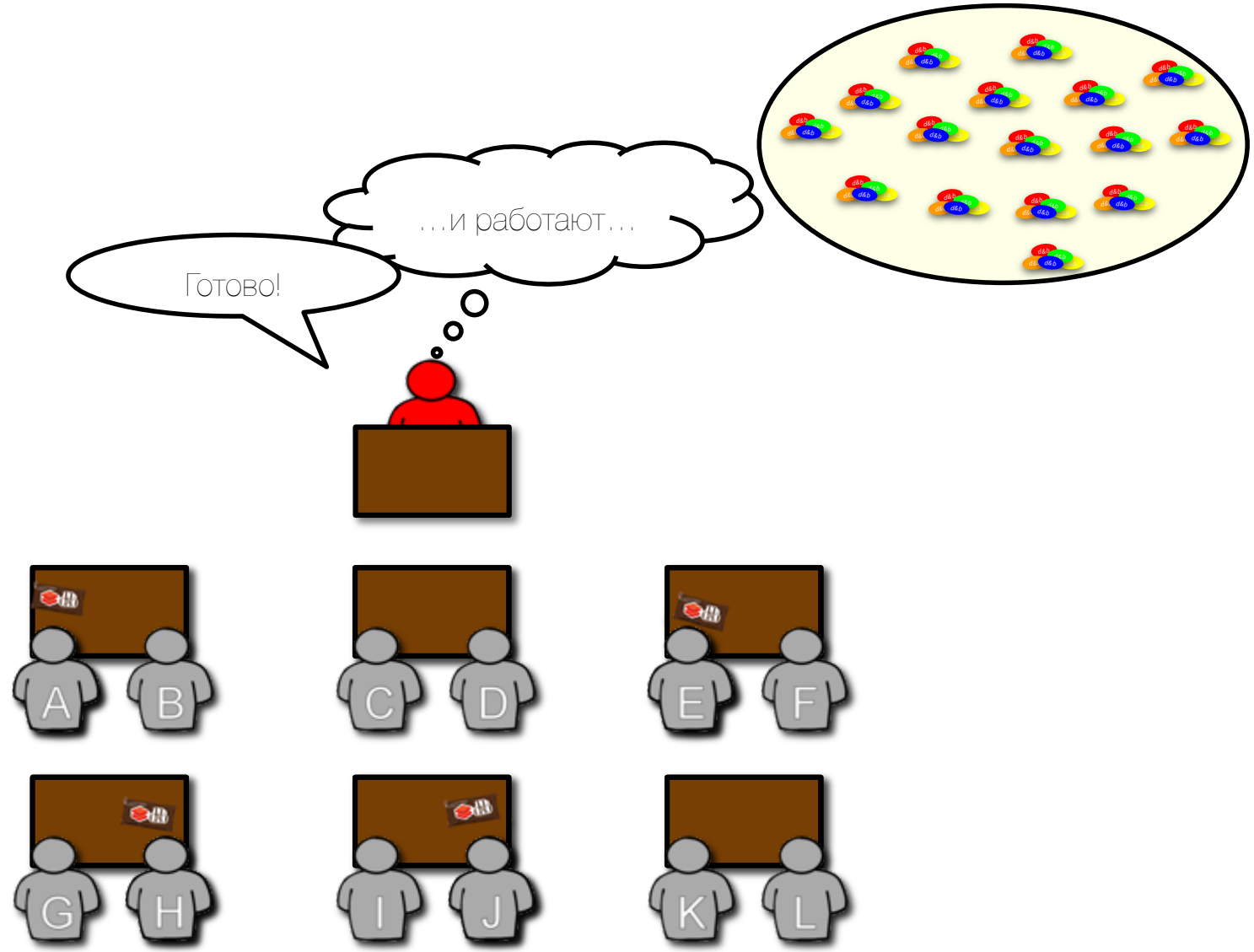

A получает пачку № 2…
B получает пачку № 3…
C получает пачку № 6…
и так далее…
A
B
C
D
E
F
G
H
I
J
K
L

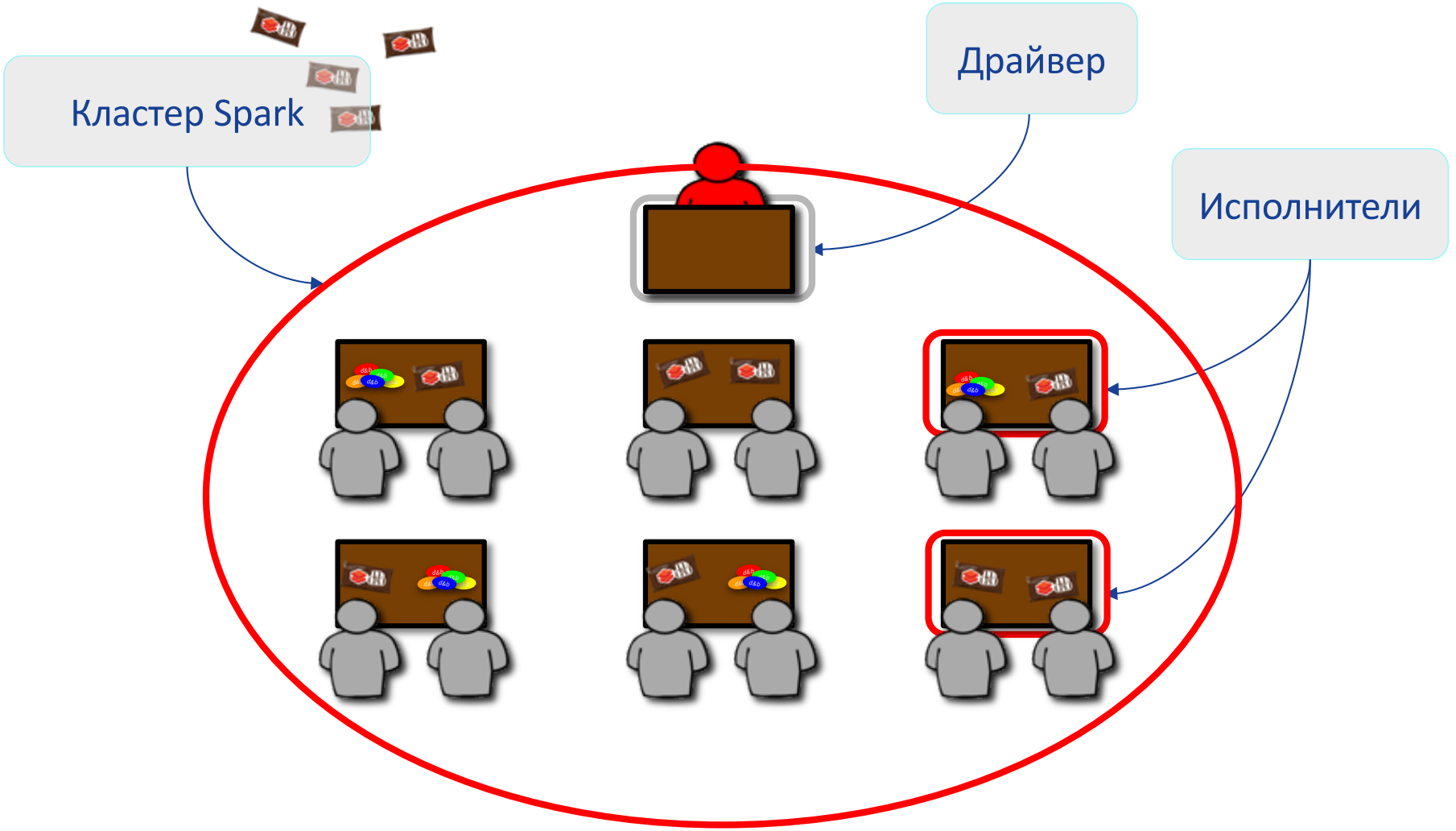
Инструкции: Съесть все коричневые и сложить остальные конфеты в кучку в углу.

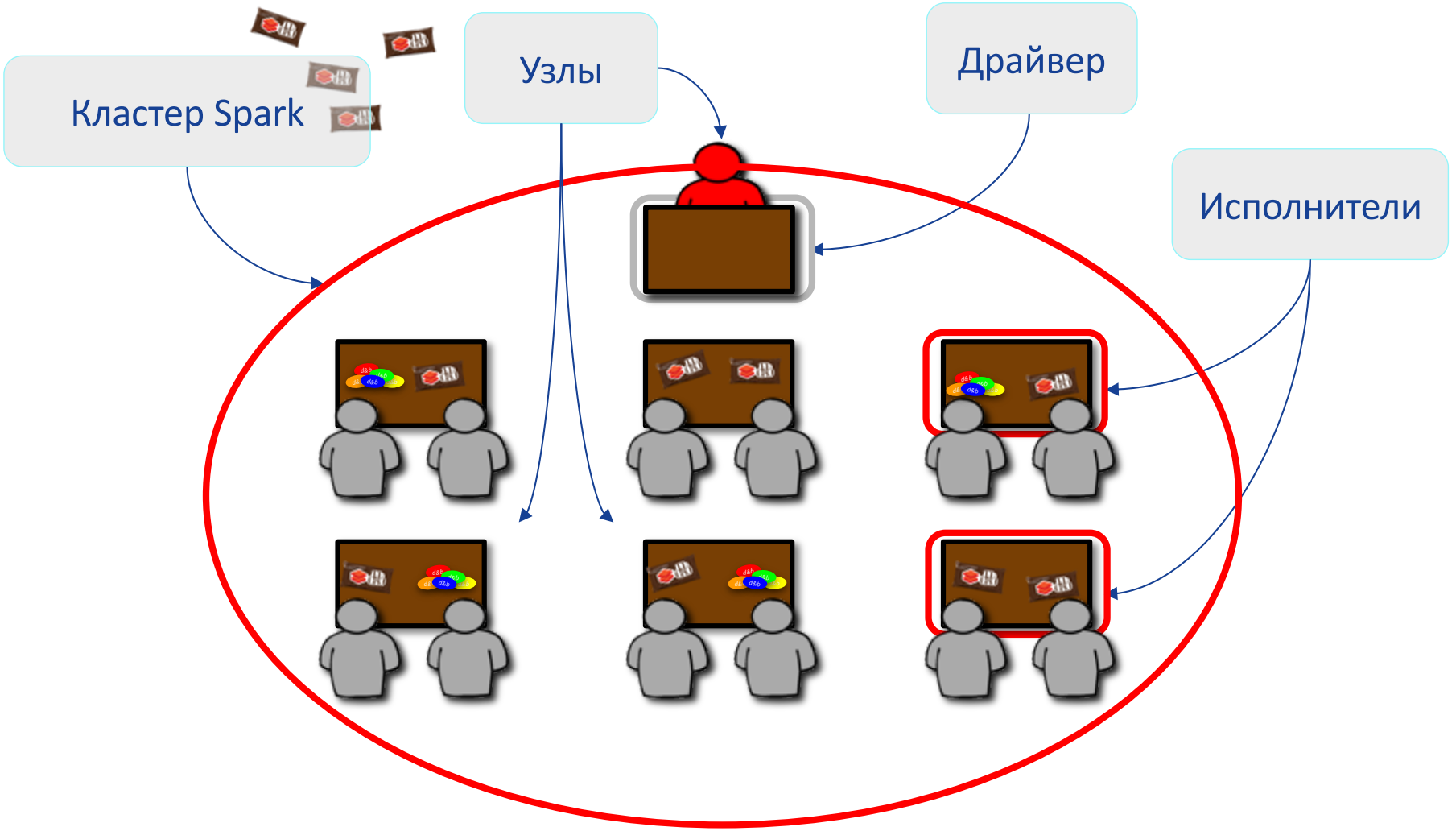


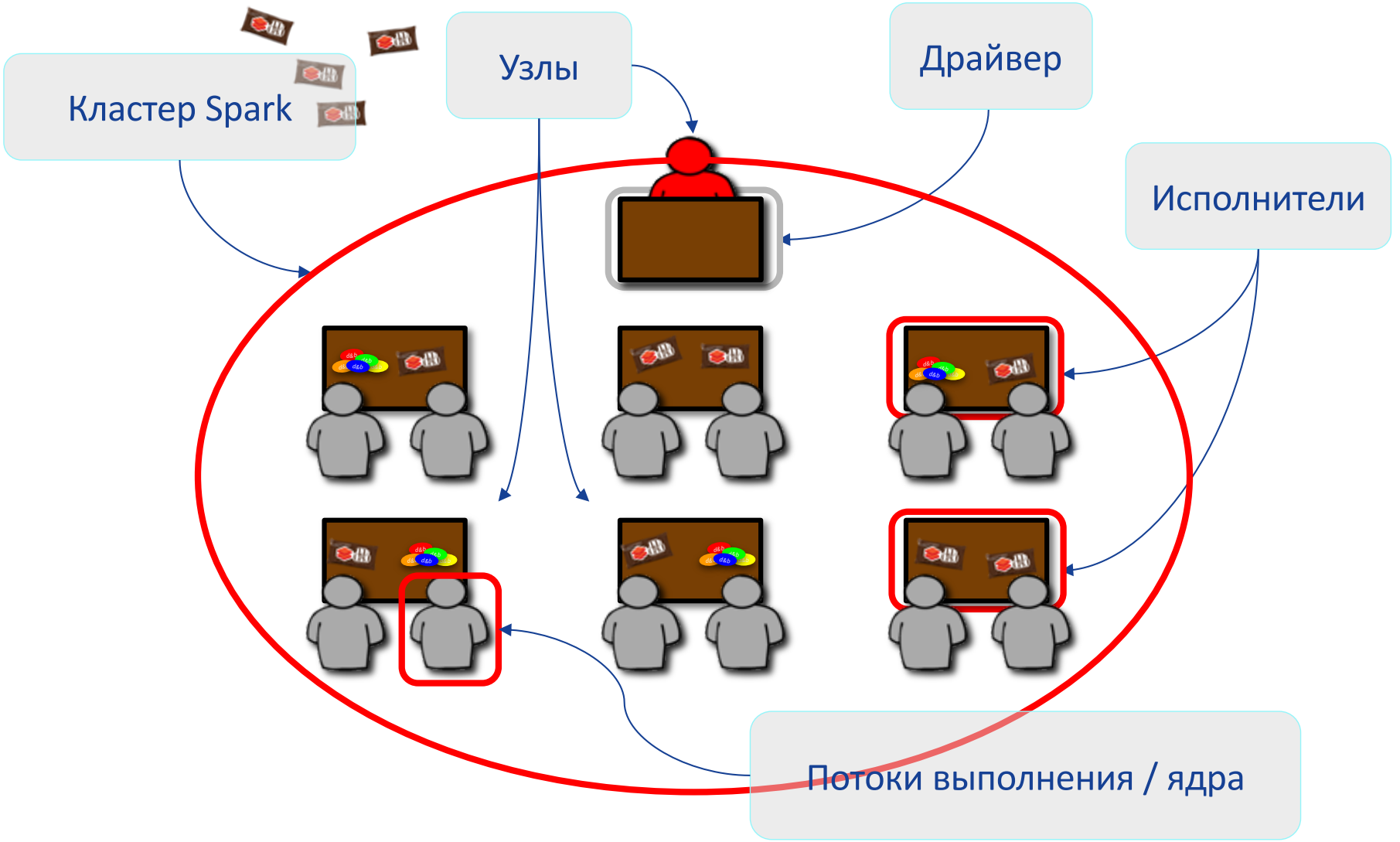

Некоторые люди быстрее других…
И в то же время другие медленнее, чем некоторые…
A
B
C
D
E
F
G
H
I
J
K
L


4 наших работника снова сидят без дела…
Им нужно дать новые задания!
A
B
C
D
E
F
G
H
I
J
K
L

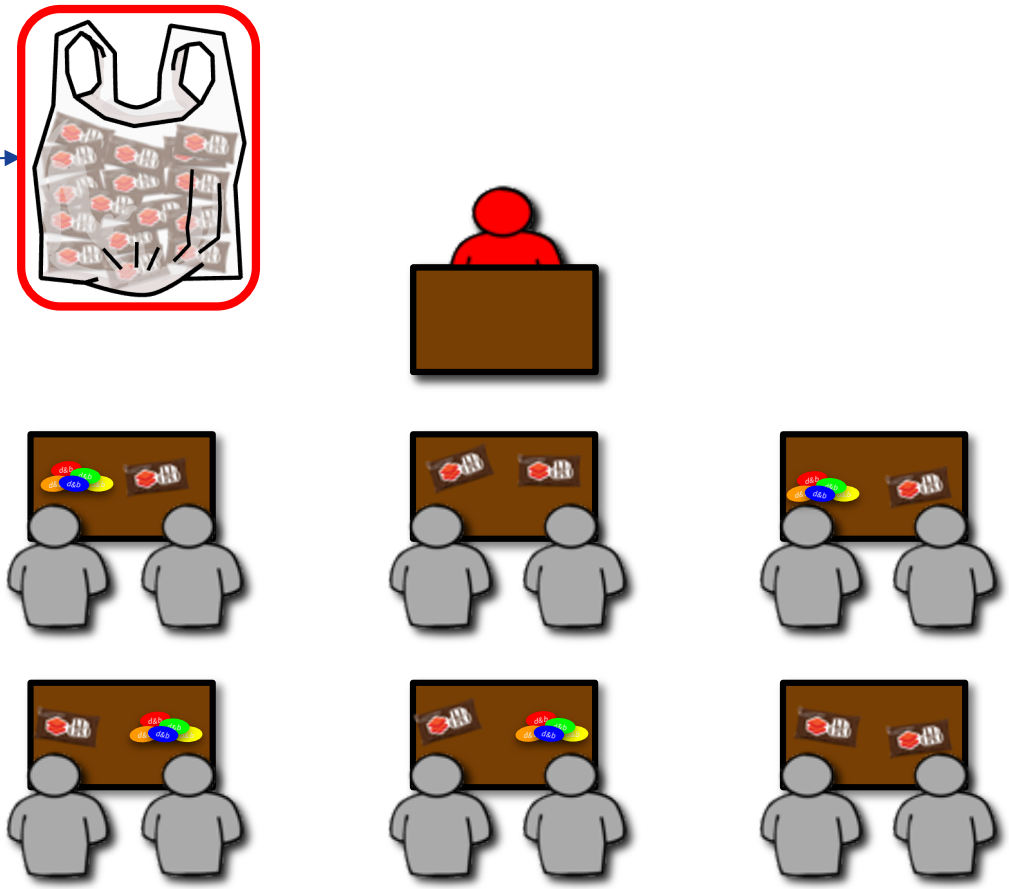

Готово!
…и работают…
A
B
C
D
E
F
G
H
I
J
K
L

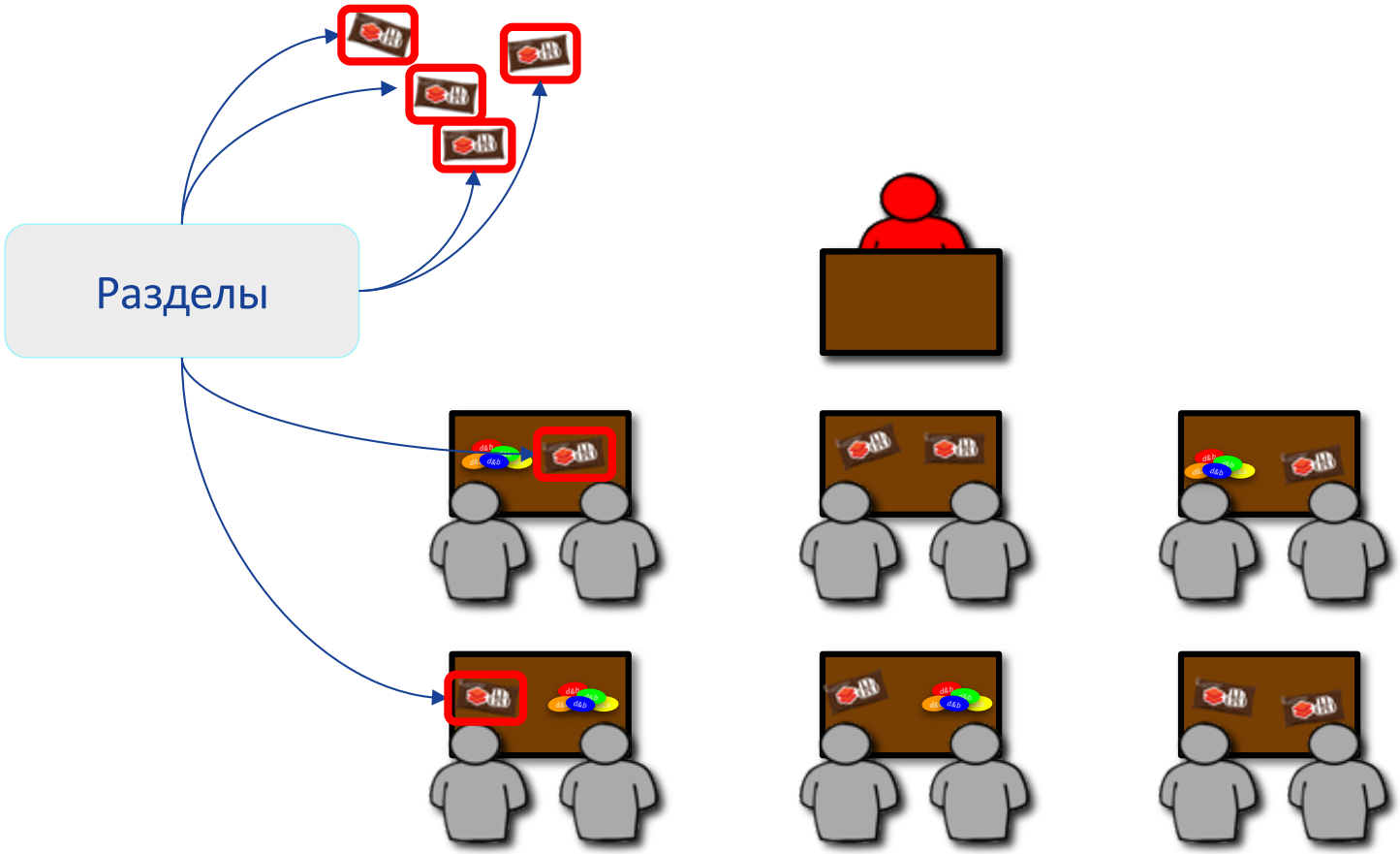

Кластер Spark
Драйвер
Исполнители

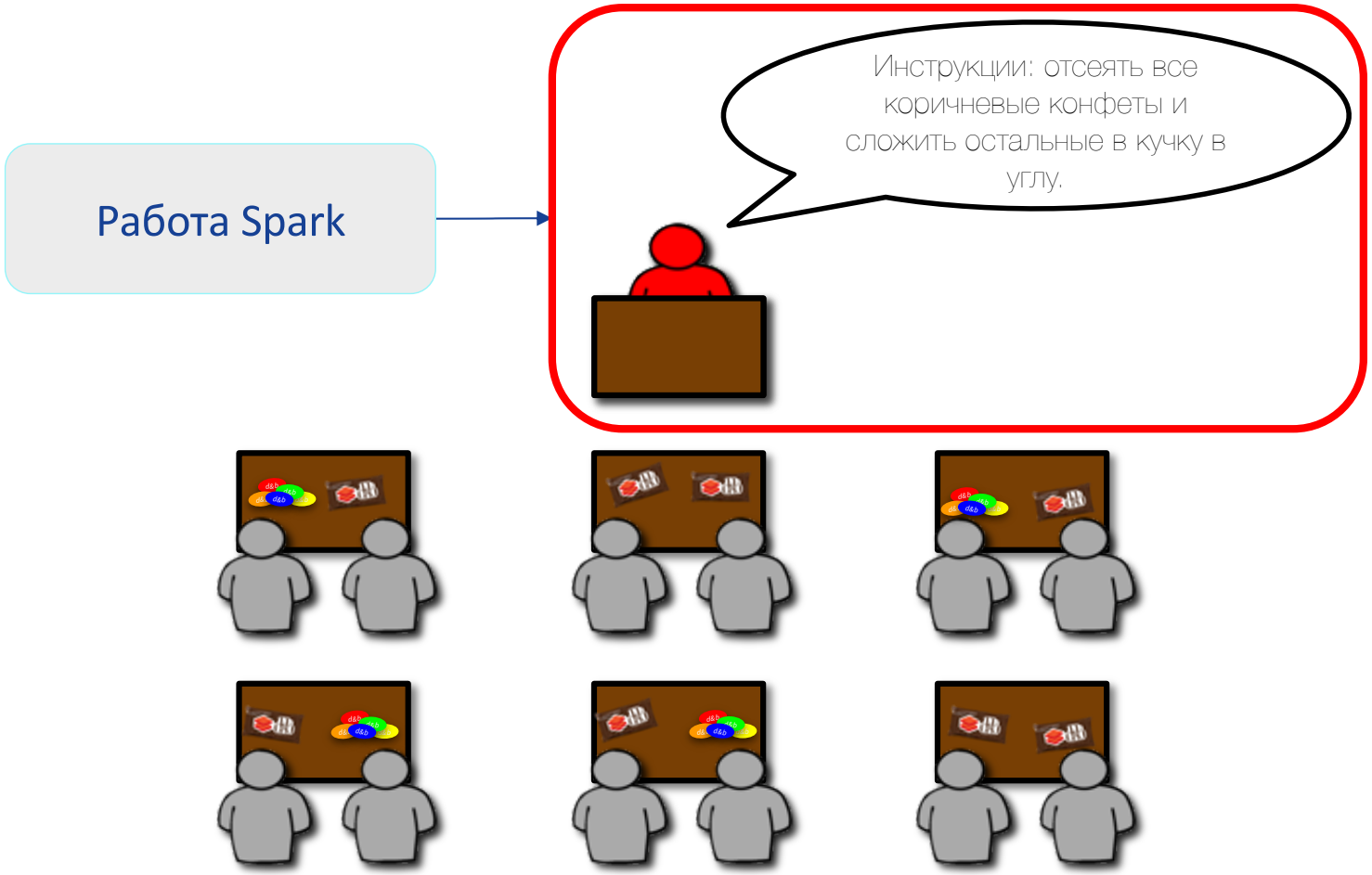

Кластер Spark
Узлы
Драйвер
Исполнители


Кластер Spark
Узлы
Драйвер
Исполнители
Потоки выполнения / ядра

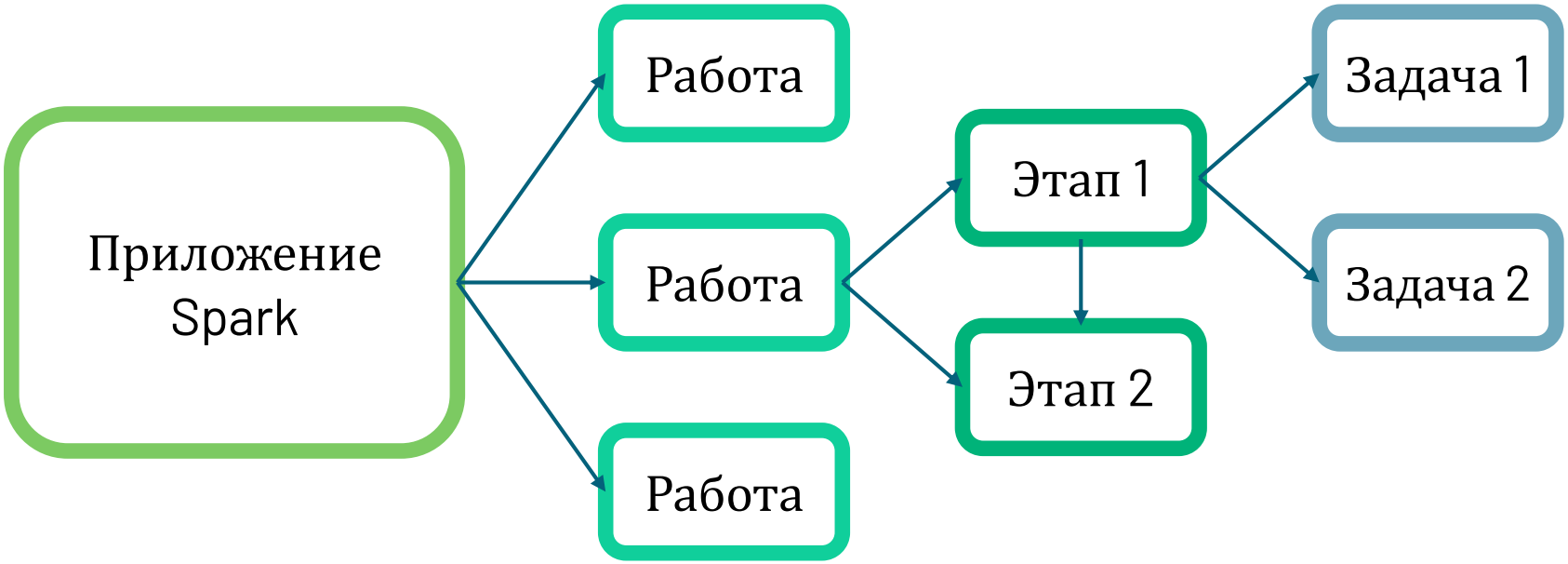
Набор данных



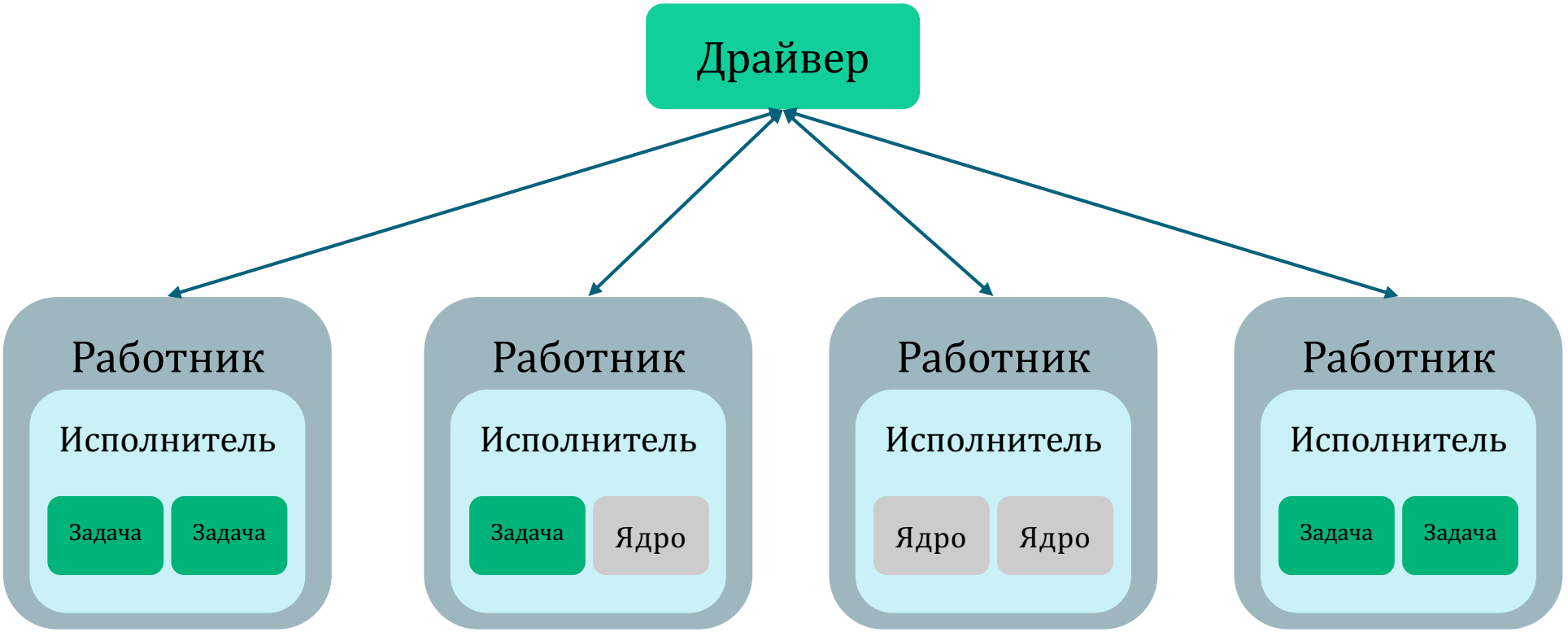

Разделы


Работа Spark
Инструкции: отсеять все коричневые конфеты и сложить остальные в кучку в углу.

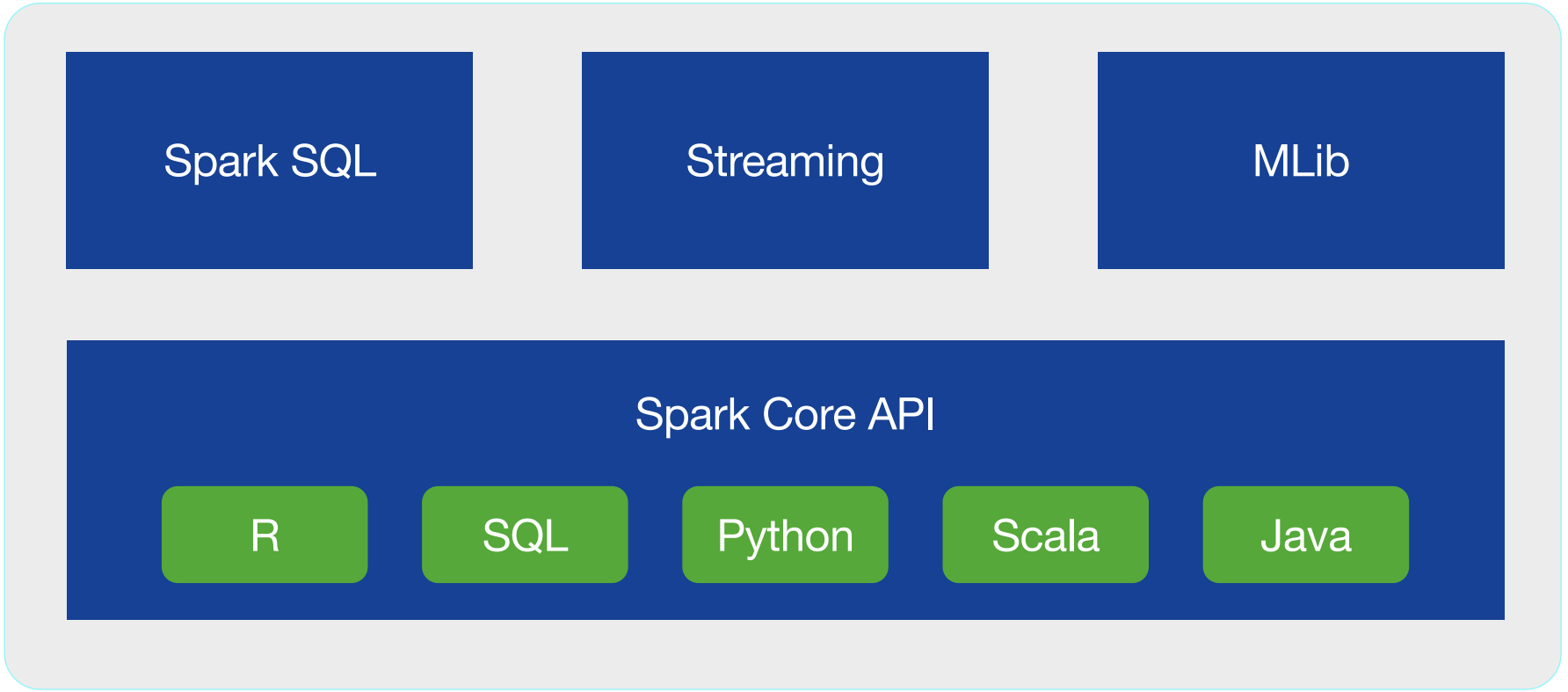
Выполнение Spark



Кластер Spark



Spark API


Spark SQL
Streaming
MLib
Spark Core API
R
SQL
Python
Scala
Java

Резюме и ключевые слова

- Spark
  - По факту стандартная обработка больших данных
- Ленивые вычисления
- Разделы набора данных
- Горизонтальное масштабирование
- Трансформации и компоненты

Вопросы?

Темы

1. Обзор и краткое повторение
2. Что такое конвейер?
3. Уровень хранения
4. Фреймворк Spark
5. **Лабораторная работа**
    1. Поиск новых наименований должностей
    2. Поиск новых профессий

Цель

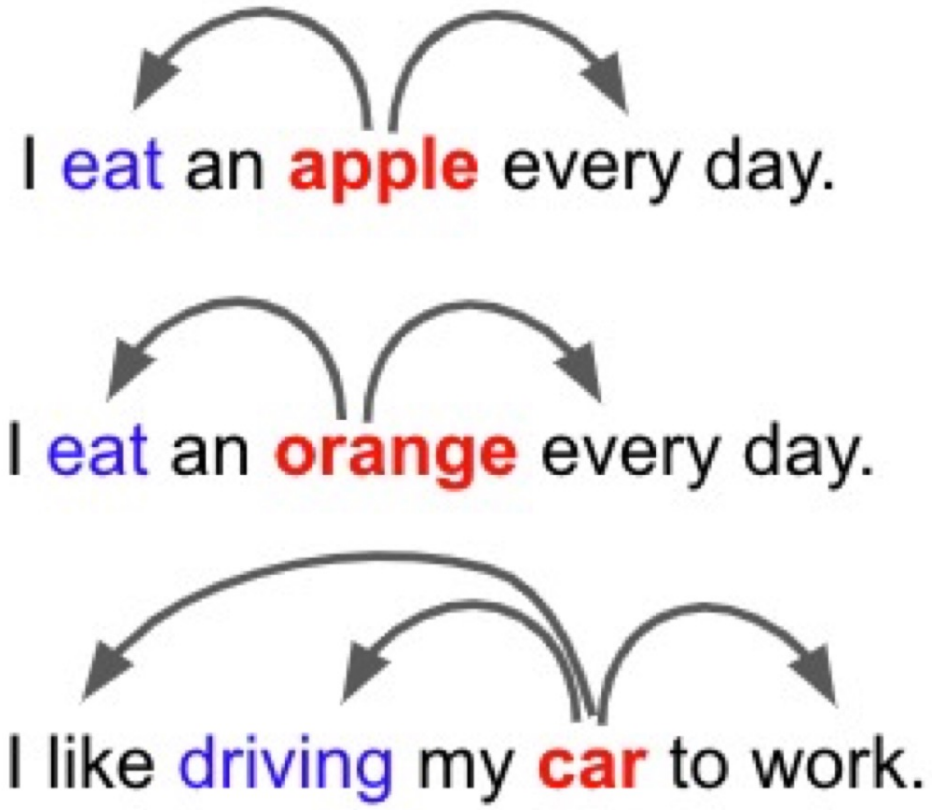
Обучить модель векторного представления слов, чтобы улучшить наши онтологии:

- начать с одной профессии
- создать корпус
- предварительная обработка
- обучить модель
- использовать модель для извлечения новых наименований вакансий

Внедрение зависит от понятия схожести слов.

Очень полезное определение — парадигматическая схожесть: Похожие слова употребляются в схожих контекстах. Они взаимозаменяемые.

Интуиция: контекст также определяет смысл



Taxonomy_improvement.ipynb

File Modifica Visualizza Inserisci Runtime Strumenti Guida Tutte le modifiche sono state salvate

Commenta Condividi

+ Codice + Testo

RAM Disco

Modifica

Taxonony improvment with Word-embeddings

Welcome!

In this notebook we first see an introduction about the concept of Word-Embedding and as we go on we'll learn how Word2Vec algorithms and see how can we implement them with the scope to improve our taxonomies (mainly ESCO occupations).

Please note that the main purpose of this notebook is to make familiar a beginner ML user with the mentioned concepts instead of focusing on the most efficient - or pythonic - way to write the code.

First we start by uploading the files we will use. This is a file with 25 observations: 5k for each occupation. We will start by processing one occupation.

```
[1] from google.colab import files
    _source = files.upload()
```

Choose Files esco_4occupations.csv

- **esco_4occupations.csv**(text/csv) - 1811573 bytes, last modified: 9/10/2020 - 100% done

```
Saving esco_4occupations.csv to esco_4occupations.csv
```

```
[2] import io
    import pandas as pd
    df = pd.read_csv(io.BytesIO(_source['esco_4occupations.csv']), sep = ',',delimiter=None, header='infer',encoding =  'utf-8')
    display(df)
```

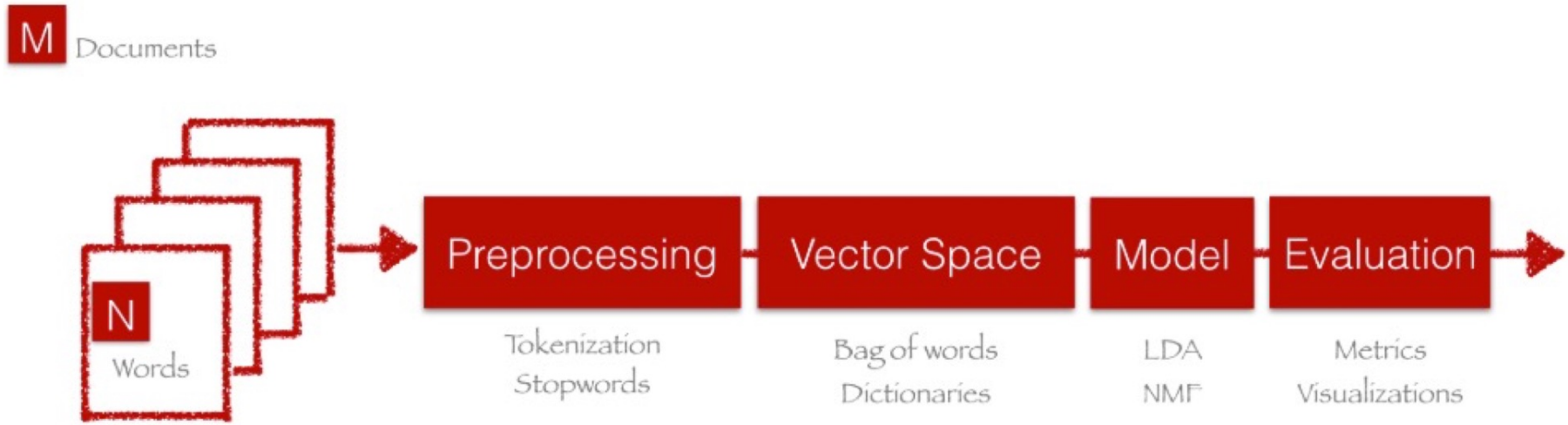
Лабораторная работа

Поиск новых профессий

Цель

Использовать модель LDA для улучшения наших онтологий и извлечения новых сведений:

- начать с корпуса вакансий
- предварительная обработка
- применить некоторые методы тематического моделирования
- извлечь новые профессии

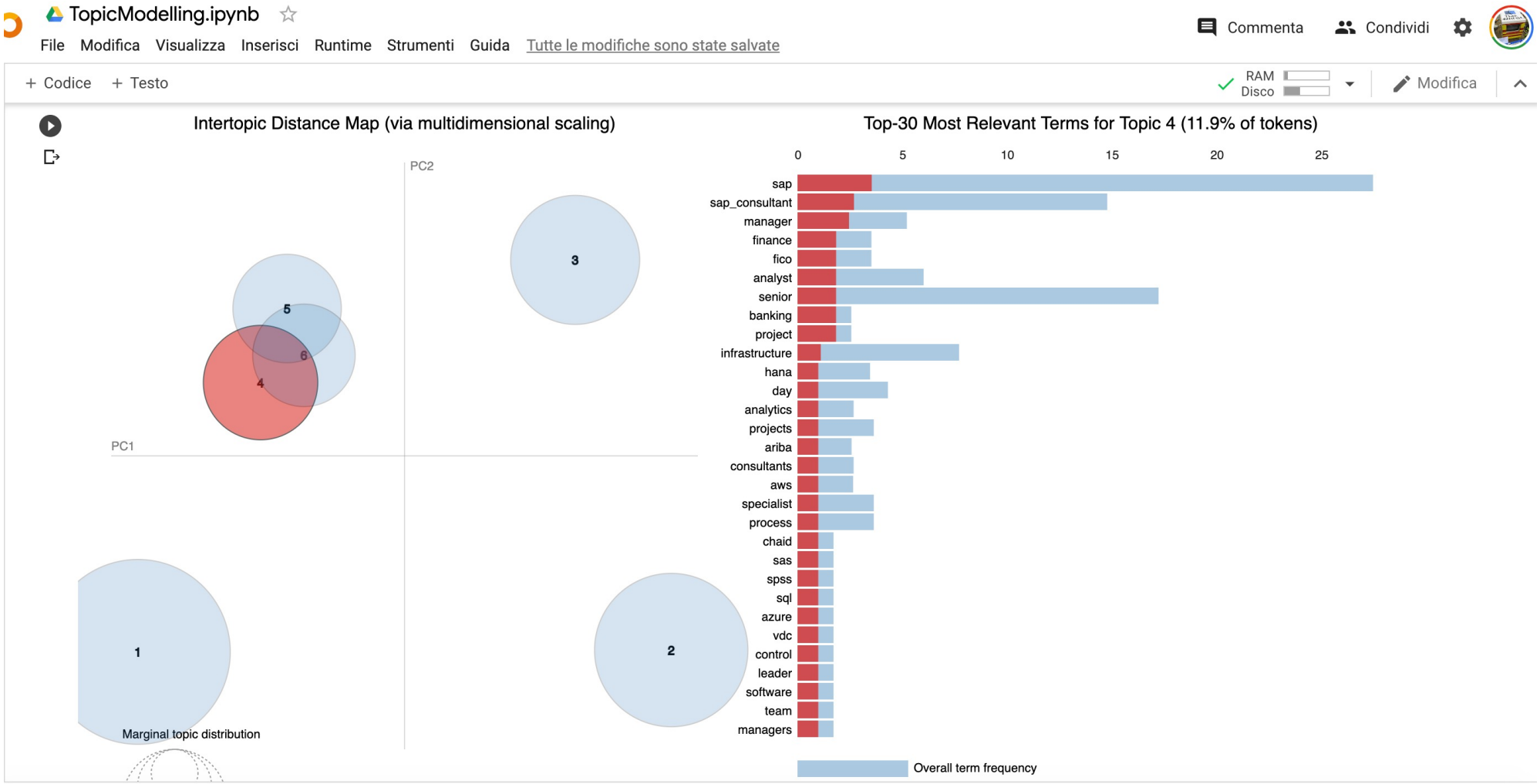

M
Documents
N
Words
Preprocessing
Vector Space
Model
Evaluation
Tokenization
Stopwords
Bag of words
Dictionaries
LDA
NMF
Metrics
Visualizations

Что означает «тема»?

Наблюдение
Одна группа слов, вероятно, будет встречаться в одном и том же **контексте**

Скрытая (т.е. неизвестная) структура, которая помогает определить, какие слова, вероятно, будут встречаться в корпусе

Тема — это распределение слов по фиксированному словарному составу


TopicModelling.ipynb
File Modifica Visualizza Inserisci Runtime Strumenti Guida Tutte le modifiche sono state salvate
Commenta
Condividi
+ Codice + Testo
RAM
Disco
Modifica
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 4 (11.9% of tokens)
0
5
10
15
20
25
sap
sap_consultant
manager
finance
fico
analyst
senior
banking
project
infrastructure
hana
day
analytics
projects
ariba
consultants
aws
specialist
process
chaid
sas
spss
sql
azure
vdc
control
leader
software
team
managers
Overall term frequency