

8.Разработване на системи с интензивни данни – Резюме

Изкуствен интелект (ИИ), машинно обучение (МО)

Data Science се занимава с **отговаряне на въпроси чрез данни**

- **Data science** включва:

- **Формиране на количествени въпроси**- могат да бъдат измерени и анализирани с помощта на данни.
- **Идентифициране на данните**, които могат да бъдат използвани за отговор на тези въпроси
- **Почистване на данните**
- **Форматиране и обработка**
- **Анализиране на данните**
- **комуникиране на резултатите** към останалите.

Big Data - различни научни области, които се занимават с извличане на стойност от големи обеми данни. Основните научни области включват: Статистика, МО, ИИ и др

Big Data Value Chain- Веригата на стойността на големите данни - процеса от събирането и обработката на данните до извлечането на полезна информация и вземането на решения на база на тези анализи.

Системи с интензивни данни (Data-intensive systems)-софтуерни и хардуерни системи, създадени за съхранение, обработка и анализ на огромни обеми от данни. **Основни характеристики:**

- 📦 **Голям обем (Volume)** – работа с терабайти или петабайти данни.
- ⚡ **Висока скорост (Velocity)** – обработка на данни в реално време или почти в реално време.

 **Разнообразие (Variety)** – данни от различни източници (текст, изображения, видео, сензори и др.).

Примери за системи с интензивни данни: Бази данни (SQL, NoSQL), Big Data платформи (Apache Hadoop, Apache Spark) – за разпределена обработка на данни; **Облачни услуги (AWS, Google Cloud, Azure)** и др.

Изграждане на система за машинно обучение (ML system) – няколко етапа, които гарантират, че моделът може да обработва данни, да разпознава модели и да прави прогнози:

1 Събиране на данни

2 Почистване и подготовка на данните

3 **Избор на модел** – използване на подходящ алгоритъм

4 **Обучение на модела** – подаване на обучаващи данни и настройка на параметрите за постигане на оптимални резултати.

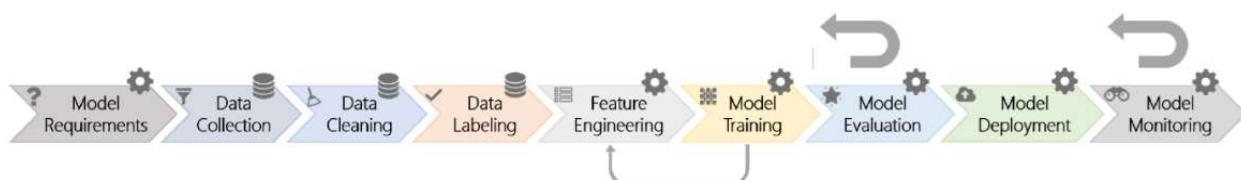
5 **Оценка на модела** – измерване на точността чрез метрики като точност (accuracy), F1-score, средна квадратична грешка (MSE).

6 **Оптимизация и настройка** – коригиране на хиперпараметрите, използване на техники като cross-validation.

7 **Деплоймънт (Разгръщане)** – интегриране на модела в реална система (уеб приложение, API, мобилно приложение).

8 **Мониторинг и поддръжка** – следене на производителността на модела, периодично актуализиране с нови данни.

Machine learning workflow, Data Science Lifecycle



Жизненият цикъл на Data Science (Data Science Lifecycle) - стъпките, които специалистите по данни следват, за да извлекат стойност от данните и да вземат информирани решения: **Формиране на проблема, Събиране на данни, Почистване и подготовка на данните, Анализ и визуализация на данните, Изграждане на модел, Оценка модела, Деплоймънт, Мониторинг и поддръжка**

DevOps - практики и автоматизацияни технологии, които подобряват **скоростта на доставяне и итерация на приложения**. DevOps позволява на Facebook да пуска десетки до стотици актуализации на всеки няколко часа.

DevOps Practices:

- **Continuous Integration (CI)**- разработчиците **автоматично изграждат, тестват и анализират** промените в софтуера при **всяка нова версия в source repository**.
- **Continuous Deployment (CD)** - инкременталните промени в софтуера автоматично се тестват, преглеждат и разгръщат в production среди.
- **Continuous Delivery (CDE)**-**гарантира, че софтуерната промяна е готова за доставка и използване от клиентите**, като се тества в production-подобни среди на тестови сървари.

AIOps (Artificial Intelligence for IT Operations) - изкуствен интелект за ИТ операции. Използва данни и машинно обучение, за подобряване и автоматизиране на операциите в информационни технологии (IT), като предоставя непрекъсната аналитична информация.

DataOps Manifesto - Values

Индивиди и взаимодействия пред процеси и инструменти
Работеща аналитика пред обширна документация

Сътрудничество с клиента пред преговори по договори

Експериментиране, итерации и обратна връзка пред обширен предварителен дизайн

Кръстосана отговорност за операциите пред изолирани отговорности

DataOps Manifesto - Principles

- **Continually satisfy your customer** – Най-важният приоритет е непрекъснатата доставка на аналитични инсайти в интервал от няколко минути до седмици.
- **Value working analytics** – Основната мярка за производителност е степента, в която аналитичните инсайти са точни, надеждни и използваеми.
- **Embrace change** – Адаптиране към променящите се клиентски нужди за конкурентно предимство.
- **It's a team sport** – Разнообразието от роли, умения и инструменти повишава иновациите и продуктивността.
- **Daily interactions** – Клиенти, аналитични екипи и operations трябва да работят заедно всеки ден по време на проекта.
- **Self-organize** – Най-добрите алгоритми, архитектури и анализи произлизат от самоорганизиращи се екипи.
- **Reduce heroism** – Целта е устойчиви и мащабириеми аналитични процеси, а не разчитане на "герои" в екипа.
- **Reflect** – Екипите трябва редовно да оценяват представянето си чрез обратна връзка от клиенти и анализи.
- **Analytics is code** – Всички аналитични инструменти генерираят код и конфигурации, описващи как данните се обработват.

- **Orchestrate** – Координацията на данни, инструменти, код и среди е ключов фактор за успеха на аналитичния процес.
- **Quality is paramount**(качество е от първостепенно значение)
 - Автоматично откриване на грешки и сигурност в кода, конфигурацията и данните.
- **Monitor quality and performance** – Постоянно наблюдение на производителността, сигурността и качеството за откриване на проблеми.
- **Reuse** – Избягване на дублиране на работа, за да се повиши ефективността на аналитичните екипи.
- **Improve cycle times** – Оптимизиране на процеса от клиентска заявка до пускане в производство и рефакторинг.

Data-Intensive Systems: С-ми, които обработват и анализират големи обеми данни, използвайки различни методологии и технологии

Основни научни области - AI, ML и Data Science: Изкуствен интелект, машинно обучение и наука за данните са основни компоненти в разработването на системи с интензивни данни.

Процеси: Включват различни етапи като събиране, съхранение, почистване и трансформация на данни.

Методологии- ориентирани към процеси; ориент. към управлен.

Process-oriented: Фокус върху автоматиз. и оптимиз. на процесите.

Management-oriented: Фокус върху управл. и координ. на екипите.

DataOps: Методология за управление на данни, която включва автоматизация и оптимизация на процесите за обработка на данни.

DataOps Manifesto: Осн. принципи и ценности, вкл. важността на взаимод. м/у хората, работещите анализи и колаборация с клиентите

MLOps: Методология за управление на машинното обучение, която включва **непрекъсната интеграция и доставка на модели**.

AIOps: Използване изк. интелект за **оптимизация на ИТ операции**

CRISP-DM (Cross-Industry Standard Process for Data Mining) – Стандартен процес за извлечение на знания от данни: Стандартен процес за разработка на проекти в областта на науката за данните. Гъвкава методология

TDSP (Team Data Science Process) е **методология, която подпомага екипите при разработката на решения, базирани на данни**. Основни компоненти на TDSP:
1/ **Роли**: Project Manager, Data Engineer, Data Scientist, Application Developer и Project Lead.
2/ **Жизнен цикъл на TDSP**: Откриване на бизнес проблема; Събиране и подготовка на данни; Изграждане на модели; Деплоймънт и поддръжка

TDSP насищава **сътрудничество, итеративен процес и автоматизация**, което прави науката за данните по-ефективна и продуктивна.

Data Driven Scrum (Управляван от данни Scrum)- адаптация на **Scrum методология** за управление на проекти, при която решенията и приоритетите се основават на данни и анализи.

DevOps: Практики и технологии за **автоматизация**, които подобряват скоростта на доставяне и итерация на приложения. **Включва**:

Continuous Integration (CI): Автоматично изграждане, тестване и анализ на софтуерни промени.

Continuous Deployment (CD): Автоматично тестване и внедряване на софтуерни промени в продукционни среди.

Continuous Delivery (CDE): Осигуряване на готовност за доставка на клиента на софтуерни промени чрез тестване в **среди, подобни на продукционните**.

Model Training и Model Inference(Обучение на модел и извеждане на модел)- 2 осн.етапа в МО:

Model Training: Процесът на обучение на модели в МО.

Model Inference: Процесът на използване на обучените модели за предсказания и анализи.

MLOps (Machine Learning Operations) е подход за **автоматизация и управление на целия жизнен цикъл на машинните модели** в производствени среди. Това включва всички етапи, от обучението на моделите, през тяхното внедряване, до тяхната поддръжка и мониторинг в реално време.

ИТ -информационни технологии

AIOps (Artificial Intelligence for IT Operations) е подход, който използва изкуствен интелект (ИИ) и машинно обучение (ML) за **подобряване на функциите на ИТ операциите** чрез предоставяне на постоянни и навременни прозорци на информация.

DataOps (Data Operations) е набор от практики, култура и технологии, които се използват за **управление на целия жизнен цикъл на данните в организация**. Основната цел на DataOps е да **автоматизира, оптимизира и подобри процесите, свързани със събирането, обработката, анализата и доставката на данни**, така че те да бъдат по-ефективни, бързи и точни. **DataOps комбинира:** Agile, DevOps, Lean Manufacturing

DataOps Manifesto

Индивиди и взаимодействия пред процеси и инструменти

Работеща аналитика пред обширна документация

Сътрудничество с клиента пред преговори по договори

Експериментиране, итерации и обратна връзка пред обширен предварителен дизайн

Кръстосана отговорност за операциите пред изолирани отговорности

Принципи: Включват удовлетворяване на клиентите, работа с аналитични данни, приемане на промени, ежедневни взаимодействия и самостоятелна организация на екипите.

Компоненти на DataOps

Оркестрация: Управление на данни, инструменти, код, среди и работа на аналитичните екипи.

Възпроизвеждане: Версиониране на данни, конфигурации и код за осигуряване на възпроизвеждани резултати.

Качество: Автоматично откриване на аномалии и проблеми със сигурността в кода, конфигурациите и данните.