

Machine Learning Spark project

Project Description

Development of data governance software based on machine learning algorithms. Working on the predicting level of confidentiality and business category for the different types of content in different languages.

Challenges

- Identification, classification, and categorization of the data in any language;
- Encrypting and protecting the data;
- Monitoring, tracking and controlling the data;
- Detection of anomalies;
- Governing the data and addressing GDPR.

CUSTOMER

NDA

INDUSTRY

Fintech
Cybersecurity

TYPE

Data science

TECHNOLOGY

Hadoop
Apache Spark
HBase
Apache Tika
Play
StanfordNLP
Deeplearning4J
Spray/Akka
Ambari

Solutions

- Application of the unsupervised, semi-supervised and supervised methods.
- Implementation of deep learning approaches to reach the high precision and process big amounts of data.
- Text summarization methods in combination with statistical solutions allowing to interpret each step.
- Using Spark Streaming to analyze data in real-time.

Summary

As a result, our experts made it possible to analyze every single electronic data file, understand its content in 70 languages and predict its level of confidentiality and business category. The targeted data-centric approach was applied in this case. The solution was to allow the platform to automatically correlate the knowledge with the user access rights to draw a clear picture.