Democratic and Popular Algeria

Ministry of Higher Education and Scientific Research
Constantine 2 University – AbdelHamidMehri

Faculty of New Information and Communication Technologies

Department of Software Technologies and Information Systems

Option : Data Science and Intelligent Systems

Comprehensive Data Preprocessing for Machine Learning Project

---

System Architecture for Cyber Attack Detection in IoT

Directed by :

Boumakh Mohamed

Goutal Wafa

Haba Doua Kamar Ezzamane

Derradji Aymene Badreddine

Under supervision :

Dr.Naila Marir

Dr.Ghammaz wafa

Dr.Benchikha Fouzia

25 avril 2024

System Architecture :

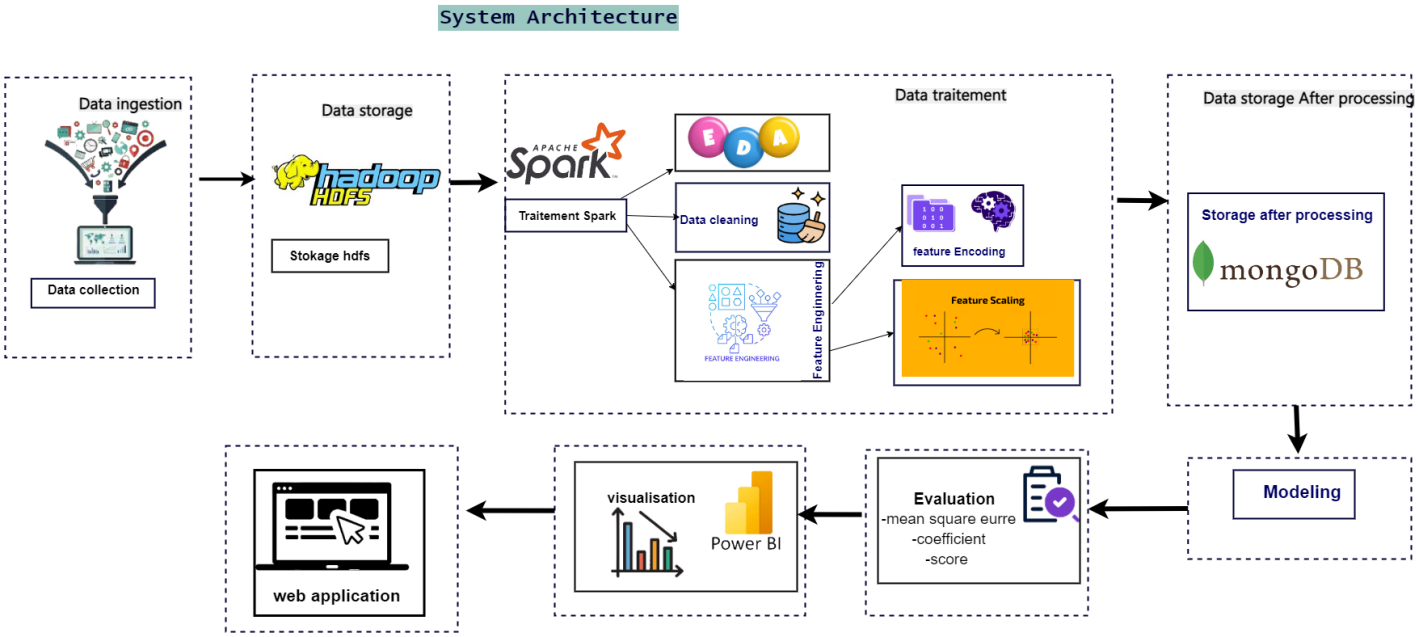

FIGURE 1 – System Architecture

1-Data Collection :

This step involves gathering the necessary data for analysis and building the machine learning model. The data can come from various sources such as databases, flat files, APIs,

2-Storage in HDFS (Hadoop Distributed File System) :

Once collected, the data is typically stored in HDFS, a distributed file system designed for storing and processing large datasets on server clusters.

3-Processing with Spark :

Apache Spark is used for large-scale data processing. The different processing steps include :

a) Exploratory Data Analysis (EDA) :

This step involves the initial exploration of the data to understand its structure, distribution, and characteristics.

(b) Data Cleaning :

Removing outliers, duplicates, missing data to ensure data quality.

(c) Feature Engineering :

Creating new features from existing data, including scaling and encoding categorical features.

4-Storage after processing in MongoDB :

Once the data has been processed, it is stored in MongoDB, a NoSQL database. MongoDB is used for its ability to handle unstructured or semi-structured data, which can be useful for storing machine learning data.

5-Modeling :

This step involves building machine learning models from the processed data. The models can include algorithms such as regression, classification, decision trees.

6-Model Evaluation :

Once the models are constructed, they are evaluated to measure their performance using various metrics such as accuracy, recall, F1 score.

7Result Visualization :

Finally, the analysis and model results are visualized for easier understanding and effective communication of the discovered insights.