CAPSTONE PROJECT SUMMARY

Data Infrastructure and Analytical Applications in Industry 4

The project's sponsor and data provider were Datapao Kft. The project's scope is factory digitization (or industry 4). The main objective is to find a solution to digitize a vegetable oil producing factory which is one of the largest in Hungary in terms of production capacity and revenue.

The research will outline the data infrastructure to collect data from the machines, its implementation, and the research activities to generate more business value for the customer.

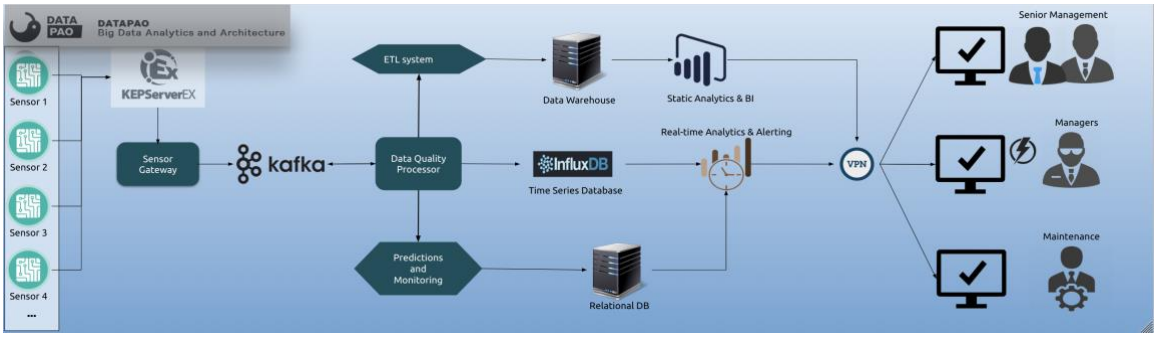
The code will be in bash, python, SQL. Databricks environment has been used for building models with Apache Spark (H2O, python and SQL API). Jupyter Notebooks and Google Colab have been used for quick prototyping.

We will make a brief introduction to Industry 4.0, its landscape and what it promises. It will be short talk about Cyber-Physical Systems (CPS), intelligent production, human-computer interaction, 3D printing, remote operations, Internet of Things, cloud computing, big data, and other modern technologies.

This story starts somewhere around February 2018, with the client approaching to Datapao Kft in order to get a proposal for **digitizing the workflows on the factory floor** and looking for opportunities to reduce utility costs like electricity in the long term. The client operates one of the largest oilseed crush plants of Central Europe using the following raw materials: Sunflower seed, rape seed, soy bean, corn germ.

First phase of the project will focus on making the data available which will be the main focus of this Capstone Project. The user interface as a BI tool will be used by the employees to get information from the machines, for making smarter business decisions.

We will talk about Designing the Solution:



The components and Python classes will be discussed briefly.

Afterwards we will list down some of the challenges:
The factory is located so far away. Internal politics of a such large organization. How we will ensure that the users will like the product and actually use it? How about data security?

In the Development chapter we will talk about forming the team and building up a digital factory product. Its architecture and how the continuous integration works will be discussed.

In the Delivery Chapter we will take a look at the actual architecture as of October 2018 and see the differences. We will list down the components of a modern big data architecture.

The chapter will be followed by more business and organizational challenges and what Datapao is doing to address the issues.

The sample data coming from the factory will be introduced right after the development details. Python notebooks covering Exploratory Data Analysis will be provided. It will be presented in the report that we would need to wait the data to accumulate to see the patterns and also, we would need labels if we want to predict machine failures (which we don't have currently) in the future.

And then the discussion will be switched to Research and Development. In this section we will develop a use case demo representing what we could achieve once the data is ready and available.

One of the possible Machine Learning applications is predicting the remaining useful life (RUL). When we have dozens of sensors collecting several measurements from all the machines in our production facility, predicting the machine failure has significant business value.

We will discuss why this will represent a big data problem and therefore Apache Spark would be used for development. The Databricks notebook will be provided which shows how to develop such applications. The solution will predict when the machine would fail.

In the final section Datapao will present these R&D activities in the international Spark Summit in London. And also, I will present it in a Spark Meetup in Cloudera Budapest office in November 2018. The links related to these events will be provided as well.

Cagdas Yetkin
November 2018, Budapest