INTRODUCTION

Over 2.5 quintillion bytes of data are created every single day and by 2020, it’s estimated that 1.7MB of data will be created every second for every person on earth [1].

This phenomenon, called **Big Data**, makes it possible to gain more complete answers because of more information and more complete answers mean more confidence in the data - which means a completely different approach to tackling problems. This tactic must manage three key features of Big Data:

* **Volume** - high volumes of low-density, unstructured data,
* **Velocity** – the fast rate at which data is received,
* **Variety** - different types of data that are available (also unstructured) [2].

In 2020 there are many technologies being used to take on this task and new ones are created each month. Behind every approach and every solution in the busines, there is a tool that enables specialists to drive a result. We can divide big data-related technologies into two groups:

* **Data Engineering** – data collection with infrastructure setting, proper formatting and generally processing it,
* **Data Analysis** – the act on the received information, getting insights into how the business works and make further decisions rely upon [3].

Both are essential in the process of creating value from received data. The goal of the thesis is to build a modern pipeline which shows how to implement ETL (extract, transform and load) data integration process with analytics workload in the Microsoft Azure cloud environment.

For the purpose of the thesis a case study involving “Wine Quality Data Set” from Machine Learning Center [4] was chosen to create a data generator to send instances of data which next can be transformed and evaluated using tools available in the Microsoft Azure cloud service:

* Azure Data Factory,
* Azure Databricks with Apache Spark and Machine Learning algorithms,
* Azure Storage – Data Lake Gen2,
* Azure SQL Server,
* Microsoft Power BI.

1 - <https://www.socialmediatoday.com/news/how-much-data-is-generated-every-minute-infographic-1/525692/>

2 - <https://www.oracle.com/big-data/guide/what-is-big-data.html>

3 - <https://mangosoft.tech/blog/top-big-data-technologies-trends-in-2019/>

4 - <https://archive.ics.uci.edu/ml/datasets/Wine+Quality>