Politechnika Warszawska

WYDZIAŁ ELEKTRONIKI
I TECHNIK INFORMACYJNYCH

Instytut Informatyki

Studia Podyplomowe
Big Data - przetwarzanie i analiza dużych zbiorów danych

PRACA KOŃCOWA

Marcin Rydelski

# Procesy ETL w środowisku Microsoft Azure

Opiekun pracy
Dr inż. Jakub Nowacki

Warszawa, 2020

Politechnika Warszawska

WYDZIAŁ ELEKTRONIKI
I TECHNIK INFORMACYJNYCH

Institute of Computer Science

Postgraduate studies

Big Data - processing and analysis of large data sets

FINAL THESIS

Marcin Rydelski

# ETL processes in Microsoft Azure environment

Thesis supervisor
Jakub Nowacki BEng, PhD

Warsaw, 2020

STRESZCZENIE

Procesy ETL w środowisku Microsoft Azure

Celem pracy jest połączenie nowoczesnych narzędzi do przetwarzania danych w celu stworzenia strumienia integracji danych w usłudze chmurowej Microsoft Azure.

Cały proces składa się z generowania, transformacji i ewaluacji danych przy użyciu technologii Big Data oraz algorytmów uczenia maszynowego. Dane wyjściowe strumienia ETL są następnie eksportowane do bazy danych Azure SQL i interaktywnego pulpitu nawigacyjnego Power BI.

Słowa kluczowe: Big Data, MS Azure, ETL, Uczenie Maszynowe, Python, Apache Spark

---

SUMMARY

ETL processes in Microsoft Azure environment

The aim of the thesis is to combine modern data processing tools to create data integration pipelines in the Microsoft Azure cloud service.

The overall process consists of data generation, transformation and evaluation using Big Data technologies and Machine Learning algorithms. The output of the ETL pipeline is next exported to Azure SQL database and interactive Power BI dashboard.

Keywords: Big Data, MS Azure, ETL, Machine Learning, Python, Apache Spark