DESTINI

SMART DATA PROCESSING AND SYSTEMS OF DEEP INSIGHT

http://www.destini2020.eu

Deliverable D6.6

Survey papers, technical reports, short papers, posters, work-in-progress papers

Document details:

Editor:	CUT
Contributors:	UNIROMA, JADS
Date:	30/09/2022
Version:	7.0

Document history:

Version	Date	Contributor	Comments
1.0	17/06/2022	A.S. Andreou S. Mappouras A. Christoforou P. Christodoulou M. Pingos	Initial document structure
2.0	20/08/2022	A. Christoforou S. Mappouras	Theses and Surveys sections added
3.0	15/09/2022	A.S. Andreou A. Christoforou	Published papers section added
4.0	25/09/2022	S. Mappouras	Work in progress section added
5.0	29/09/2022	A.S. Andreou	Final Review
6.0	30/09/2022	Partners	Partners Review
7.0	20/12/2022	A.S. Andreou	Document Revision

Contents

Contents

1. Introduction

1.1 Purpose

This document includes all outcomes of the project in terms of survey papers, technical reports, short papers, posters, and work-in-progress papers that were produced through its activities.

This deliverable is part of Workpackage-6 (WP6) produces and carries out the dissemination and communication strategy of the project. It also develops a staged approach to how the work has been conducted, published and shared.

1.2 Definitions, Acronyms, and Abbreviations

CUT: Cyprus University of Technology

UNIROMA: Sapienza University of Rome

ERRIS/JADS: European Research Institute in Service Science / Jheronymus Academy of Data Science

1.3 Overview

The rest of the document is structured based on the research work the project partners have conducted through the project's activities, categorized by type. Section 2 presents undergraduate and postgraduate theses. Section 3 describes papers published thus far. Section 4 includes surveys and section 5 works in progress. Finally, section 6 concludes the document regarding the work presented along with the research work that is currently active.

2. Theses

In the context of the DESTINI project, a series of undergraduate and postgraduate theses have been assigned to students from CUT. The content of these theses was fully in line with the topics and research pillars of the project. The students that undertook these theses have actively participated in project activities organized by the consortium, mainly summer schools and workshops.

1.4 Undergraduate Theses

BPM in Healthcare

Business process mining is a must-have set of techniques for top management to better organize and automate operational processes. Process management enables business owners to turn processes into visual flows and flows into automations. This is also the way to keep operations aligned with goals and strategies, track performance, and detect gaps or process bottlenecks to fix. This thesis aims to apply BPM models, specifically process discovery, conformance checking and process optimization, to Healthcare data. The thesis utilizes data from health services in Cyprus (e.g., Ambulance).

BPM on Data Lakes

Process mining is a family of techniques combining data science and process management to support the analysis of operational processes based on event logs. Process mining aims to turn event data into insights and actions. Process mining techniques use event data to show what people, machines, and organizations are really doing. Process mining provides novel insights that can be used to identify the execution path taken by operational processes and address their performance and compliance problems. This thesis's goal is to apply Business Process Mining (BPM) by processing not only event logs (structured data) but also semi-structured and unstructured data stored in their raw/natural format in Data Lakes system repositories. The thesis experimented with data that are expressed as RDBMS tuples, images, videos, tweets etc. The data are first transformed into a standardized format (e.g. metadata) and then BPM techniques are applied to this format.

Blockchain - Data Lakes Blueprint - Metaverse – NFTs

The Metaverse is an online, three-dimensional universe that combines multiple virtual spaces. It can be compared to a future version of the internet. With Metaverse, users can collaborate, meet, play games, and socialize in these 3D spaces. NFTs are digital items that can be bought and sold using this blockchain technology. Users can have complete control over their digital assets in the Metaverse, thanks to NFTs. Blockchain technology provides immutable confirmation of ownership that underpins these virtual worlds. This thesis aims to use Blockchain and NFTs' technologies to make Data Lakes Blueprint (an existing work) available to the Metaverse.

Methodology for assessing the accuracy of logs in Process Mining

This thesis investigates the challenge of the accuracy of logs used in the area of Process Mining to perform process discovery. Data accuracy is paramount for producing a reliable process model and then experimenting with it to improve tasks and steps within the discovered process. If accuracy is low, then any decision made to modify and optimize a process or any information retrieved from the analysis of the data projects on this model is biased and flawed. Therefore, the thesis surveys the literature to find appropriate metrics and techniques that could enable the assessment of the accuracy level of logs and proposes possible ways to improve it.

1.5 Master Theses

Real-time processing and visualization of heterogeneous data streams

The purpose of this thesis is a methodological approach to gather data streams and discover data values. All this information is taken and represented with Digital Twin. Digital Twin monitors the structure of streaming data to interact with gathering data to customize and finally visualizes the data with graphical techniques.

Graphical techniques to show how people and tasks are integrated into reality

The purpose of this thesis is to collect all the information of task flow that people execute a business process from event data and process logs. This is achived with the Digital Twin to define and collect the flow of tasks and metrics (total time of execution, number of steps, time for each step, etc). Finally, all this information is shown in a graphical dashboard that studies business flows.

Interactive Dashboard for business flow changes using Blockchain

This thesis aims to interact with the business flow to reduce time, cost and human resources. This interaction must record all the flow and parameter changes in Blockchain. Moreover, the Blockchain gives the resources and analytical changes to users. When a business flow succeeds in reducing cost, for example, updating all the users using Blockchain technology.

3D Training platform

This thesis aims to create a 3D training platform to train and increase employees' skills for resolving real problems in production line machines. The 3D environment is being designed in the virtual world (Metaverse) with if-scenarios and interactions. All steps and processes of training are stored on Blockchain to certify the training.

Business Process Mining with Visual Querying

This master thesis uses Digital Twin to mainly use a graphical technique to investigate the relations between the data. Users can use standard steps to obtain the desired result without programming skills. This thesis converts process mining into an interactive procedure that utilizes Digital Twins to visualize data of historical data and processes that were retrieved from logs or data warehouses.

Metaverse in Healthcare

This thesis aims to create a virtual environment that visualizes patients' files. Users can read and interact with humans in the virtual environment to see the patient's problem and suggest treatment. The idea of this interaction patient is to propose a treatment and see the progress of treatment in real-time.

3. Surveys

Survey on Industry 4.0 - Smart Data Processing and Systems of Deep Insight Current Research and Future Challenges

This survey is conducted in the context of a DESTINI project which aims to identify and quote the most significant research findings, challenges and open problems on Smart Data Processing and Systems of Deep Insight approaches in the area of Industry 4.0 and Smart manufacturing that are reported in the relevant literature. The survey is organized as follows: First, the survey presents the research questions that motivate this study and describes the methodology followed to identify relevant studies published in various venues. Secondly, it outlines the most important aspects of these studies organized in specific scientific areas accompanied by their literature review, introducing the problem dealt with, the methodology followed, and the results produced. The scientific areas included are: Infrastructures, frameworks and technologies supporting SDP (Data Lakes, Data Meshes, CPS) and Tools and Techniques supporting SDI (Predictive Maintenance and predictive analytics, BPM, Blockchain, Optimization, Decision Support and Prediction) in the area of Smart Manufacturing. Finally, the survey summarizes the research challenges and open problems identified in the corresponding studies reviewed.

Survey on Graphical methods and models that contribute to the area of Smart Data to identify the most significant challenges and open problems

The new scientific trends nowadays worldwide are the Internet of things (IoT), big data, cloud computing, artificial intelligence (AI) and other new generation information technologies. All these generate a large volume of data that may be structured, semi-structured and unstructured. Big data analysis models and algorithms can run to organize, analyze and mine these raw data to obtain valuable knowledge. These data, when visualized you, can provide different information with the use of some filters. Data visualization represents data in some systematic form, including attributes and variables for the unit of information. Visualization data allows users and businesses to mash data sources to create custom analytical views.

In manufacturing, we see that big data involve a large volume of structured, semi-structured and unstructured data generated from the product lifecycle. Internet of Things (IoT) devices collect these manufacturing data in real time and sometimes automatically. Manufacturers aim to find a way to increase efficiency, manage the storage of all these data and visualize them to improve and increase productivity and quality of manufacturing. Nowadays, industries are being transformed with the rise of IoT, autonomous robots, cyber-physical systems, cloud computing and cognitive computing. This transformation is called Industry 4.0 or Smart Industry. Industry 4.0 aims to construct an open, smart manufacturing platform for industrial information applications based on various technologies.

This survey aims to investigate specific approaches within the above areas. Furthermore, the study will involve Large-scale data analytics, decision monitoring and next best action, and descriptive, predictive and cognitive analytics.

4. Published Papers

1. skillsChain: A Decentralized Application That Uses Educational Robotics and Blockchain to Disrupt the Educational Process

Panayiotis Christodoulou, Andreas S Andreou, Zinon Zinonos

DOI: 10.3390/s21186227

Abstract

Our epoch is continuously disrupted by the rapid technological advances in various scientific domains that aim to drive forward the Fourth Industrial Revolution. This disruption resulted in the introduction of fields that present advanced ways to train students as well as ways to secure the exchange of data and guarantee the integrity of those data. In this paper, a decentralized application (dApp), namely skillsChain, is introduced that utilizes Blockchain in educational robotics to securely track the development of students' skills so as to be transferable beyond the confines of the academic world. This work outlines a state-of-the-art architecture in which educational robotics can directly execute transactions on a public ledger when certain requirements are met without the need of educators. In addition, it allows students to safely exchange their skills' records with third parties. The proposed application was designed and deployed on a public distributed ledger and the final results present its efficacy.

2. KnowGo: An Adaptive Learning-Based Multi-model Framework for Dynamic Automotive Risk Assessment

Paul Mundt, Indika Kumara, Willem-Jan Van Den Heuvel, Damian Andrew Tamburri & Andreas S. Andreou

(https://doi.org/10.1007/978-3-031-11510-3_18), Part of the Lecture Notes in Business Information Processing book series (LNBIP,volume 453)

Abstract

In autonomous driving systems, the level of monitoring and control expected from the vehicle and the driver change in accordance with the level of automation, creating a dynamic risk environment where risks change according to the level of automation. Moreover, the input data and their essential features for a given risk model can also be inconsistent, heterogeneous, and volatile. Therefore, risk assessment systems must adapt to changes in the automation level and input data content to ensure that both the risk criteria and weighting reflect the actual system state, which can change at any time. This paper introduces KnowGo, a learning-based dynamic risk assessment framework that provides a risk prediction architecture that can be dynamically reconfigured in terms of risk criterion, risk model selection, and weighting in response to dynamic changes in the operational environment. We validated the KnowGo framework with five types of risk scoring models implemented using data-driven and rule-based methods.

3. A Data Lake Metadata Enrichment Mechanism via Semantic Blueprints

Michalis Pingos, Andreas S. Andreou

Evaluation of Novel Approaches to Software Engineering (ENASE), 2022

DOI: 10.5220/00110 80400003176

Abstract

One of the greatest challenges in Smart Big Data Processing nowadays revolves around handling multiple heterogeneous data sources that produce massive amounts of structured, semi-structured and unstructured data through Data Lakes. The latter requires a disciplined approach to collect, store and retrieve/ analyse data to enable efficient predictive and prescriptive modelling, as well as the development of other advanced analytics applications on top of it. The present paper addresses this highly complex problem and proposes a novel standardization framework that combines mainly the 5Vs Big Data characteristics, blueprint ontologies and Data Lakes with ponds architecture, to offer a metadata semantic enrichment mechanism that enables fast storing to and efficient retrieval from a Data Lake. The proposed mechanism is compared qualitatively against existing metadata systems using a set of functional characteristics or properties, with the results indicating that it is indeed a promising approach.

4. Unsupervised Labor Intelligence Systems: A Detection Approach and Its Evaluation. A Case Study in the Netherlands.

Giuseppe Cascavilla, Gemma Catolino, Fabio Palomba, Andreas S. Andreou, Damian A. Tamburri & Willem-Jan Van Den Heuvel

Conference paper First Online: 01 October 2022, Part of the Communications in Computer and Information Science book series (CCIS,volume 1603)
DOI: 10.1007/978-3- 031-18304-1_5

Abstract

In recent years, job advertisements through the web or social media represent an easy way to spread this information. However, social media are often a dangerous showcase of possibly labor exploitation advertisements. This paper aims to determine the potential indicators of labor exploitation for unskilled jobs offered in the Netherlands. Specifically, we exploited topic modeling to extract and handle information from textual data about job advertisements for analyzing deceptive and characterizing features. Finally, we use these features to investigate whether automated machine learning methods can predict the risk of labor exploitation by looking at salary discrepancies. The results suggest that features need to be carefully monitored, e.g., hours. Finally, our results showed encouraging results, i.e., F1-Score 61%, thus meaning that Data Science methods and Artificial Intelligence approaches can be used to detect labor exploitation—starting from job advertisements—based on the discrepancy of delta salary, possibly representing a revolutionary step.

5. DLMetaChain: An IoT Data Lake Architecture Based on the Blockchain

Michalis Pingos, Panayiotis Christodoulou, Andreas S. Andreou

13th International Conference on Information, Intelligence, Systems & Applications (IISA)

DOI: 10.1109/iisa56 318.2022.9904 404

Abstract

Nowadays, the IoT ecosystem is evolving rapidly, with multiple heterogeneous sources producing high volumes of data and processes transforming this data into meaningful or “smart” information . These volumes of data, including IoT data, need to be stored in repositories that can host raw, unprocessed, relational and non-relational types of data, such as Data Lakes. Due to the weakness of metadata management, security & access control is one of the main challenges of Big Data storage architectures as Data Lakes can be replaced without oversight of the contents. Recently, the Blockchain technology has been introduced as an effective solution to build trust between different entities, where trust is either nonexistent or unproven, and to address security and privacy concerns. In this paper we introduce DLMetaChain, an extended Data Lake metadata mechanism that consists of data from heterogeneous data sources which interact with IoT data. The extended mechanism mainly focuses on developing an architecture to ensure that the data in the Data Lake is not modified or altered by taking into advantage the capabilities of the Blockchain.

6. A Smart Manufacturing Data Lake Metadata Framework for Process Mining

Michalis Pingos, Andreas S. Andreou

ICSEA 2022, The Seventeenth International Conference on Software Engineering Advances

DOI: N/A

Abstract

The fourth industrial revolution consists of a new level of organization and control of the entire production process. Smart manufacturing ecosystem and especially Cyber Physical Systems are evolving rapidly. They constitute an environment with multiple heterogeneous sources that produce high volumes of data. This data need to be stored in a storage system that can handle raw, unprocessed, relational, and non-relational data types, such as Data Lakes, in order to be processed when needed and bring insight. This paper introduces a Data Lake-based metadata framework, which utilizes the concept of blueprints to characterize the data sources and the data itself to facilitate process mining tasks. The applicability and effectiveness of the proposed framework is validated through a real-world smart manufacturing case-study, namely a poultry meat production factory, which offers operational support and business workflow analysis.

7. An Interactive Digital Twin for Visual Querying and Process Mining

Spyros Loizou, Andreas S. Andreou

ICSEA 2022, The Seventeenth International Conference on Software Engineering Advances

*DOI: N/A***Abstract**

Large volumes of structured, semi-structured and unstructured data are produced daily by industrial businesses which require analysis and processing with appropriate models and algorithms to obtain valuable knowledge. This paper introduces a framework for business technology that combines the notion of Digital Twins with Process Mining aiming at delivering a simple and efficient way to retrieve customized data and process it with the use of graphical techniques, providing interactive visualization of process mining steps. More specifically, the proposed framework provides the ability to define different data sources and link these sources with a visual query generator which constructs, executes and depicts graphically the results of custom queries. The framework includes also sophisticated Artificial Intelligence / Machine Learning algorithms for data analysis, filtering and prediction. The framework is demonstrated through an interactive dashboard, which was implemented in Python to support a fully operational and visual process mining environment that facilitates decision making without the need of programming or data management skills.

8. Exploiting Metadata Semantics in Data Lakes Using Blueprints

Michalis Pingos,Andreas S. Andreou

Springer Book, not acceptet yet.

Abstract

Abstract. Smart processing of Big Data has been recently emerged as a field that provides quite a few challenges related to how multiple heterogeneous data sources that produce massive amounts of structured, semi-structured and unstructured data may be handled. One solution to this problem is manage this fusion of disparate data sources through Data Lakes. The latter, though, suffers from the lack of a disciplined approach to collect, store and retrieve data to support predictive and prescriptive analytics. This chapter tackles this challenge by introducing a novel standardization framework for managing data in Data Lakes that combines mainly the 5Vs Big Data characteristics and blueprint ontologies. It organizes a Data Lake using a ponds architecture and describes a metadata semantic enrichment mechanism that enables fast storing to and efficient retrieval. The mechanism supports Visual Querying and offers increased security via Blockchain and Non-Fungible Tokens. The proposed approach is compared against other known metadata systems utilizing a set of functional properties with very encouraging results.

5. Work in progress

Fuzzy Cognitive Maps (FCMs): Identification, correlations and matching in the Forensic DNA profiling

The aim of that collaboration is to go through the phases that involve the police, the analysis center of the forensic laboratories to help and support the last and decisive phase of result's interpretation. The problems to assign the right profile to the murderer, and the people involved in the crime scene, still critical. DNA profiling is a procedure that can be used to identify individuals on the basis of their unique DNA to solve crime and investigations. Forensics analysis focuses on allele recognition and the individuation of repeating sequences in human DNA. During the DESTINI meeting, we discussed and evaluated an alternative approach to support the actual and manual interpretation of results from lab analysis. The difficulty in interpretation depends on abnormalities occurring in the laboratory during the genome amplification process (PCR). We have discussed and defined the future steps in that project:(1) Exploring the Forensics Data of Forensics Institute, (2) making Reseach Questions in line with the problem and more specific,(3) Starting with simple algorithms to extract significative informations, (4) analysing previous results and critical aspects came out during the PDEng program of JADS, (5) define an Architecture (6) Looking to find correlations with Fuzzy Cognitive Maps (FCMs) model and define how proceed in the future for this Forensics Project and collaboration with others Forensics groups.

Usage of FCM for evaluating a Maturity Model

Industry 4.0 has introduced many technologies requiring absolute knowledge to be implemented correctly. Because of the multitude of solutions and techniques, it is not easy for a company to schedule and plan the roadmap and the investments required to shift towards Industry 4.0. Moreover, it is not straightforward for a company to understand its readiness as an Industry 4.0 player. The presence of a maturity model to assess the maturity and readiness of a company as an Industry 4.0 actor, according to the complexity and the type of software and hardware installed and their usage, can bring significant advantages. Companies can use it to evaluate and analyze their strengths and weaknesses, but, more importantly, they can use it to define a roadmap of investments to reach a higher "maturity level." This work aims to provide a maturity model based on Multi-Layer Fuzzy Cognitive Map approach.

6. Conclusions

This deliverable is part of Work-Package 6 (WP6) which develops the dissemination and communication strategy of the project. More specifically, this deliverable presents the preliminary results of the research steps that the partners facilitate through closed collaboration activities. Survey papers, student theses, papers and work-in-progress (end of the project) are prepared and presented, aiming to collect valuable feedback from researchers and experts in the fields of study, which will be valuable for coordinating and directing future research activities