Dark Data in Internet of Things (IoT): Challenges and Opportunities

Dimitar Trajanov, Vladimir Zdraveski, Riste Stojanov and Ljupco Kocarev

Abstract - Nowadays we are witnessing the establishment of the data-driven science as a new scientific paradigm, that is opening a waste amount of new opportunities for scientific and technological advances. The data is becoming the main asset in today's science and technology. Unfortunately, a significant amount of available and stored data is not used today. This data is known as a dark data. Starting from this point, the primary goal of this paper is to raise the awareness of the opportunities that are explored with the dark data utilization in companies and organizations, by giving an overview of the underlining technologies, proposing a methodology and showing example projects that utilize the dark data in the IoT domain.

Keywords – Dark data, Internet of Things (IoT), Machine Learning, Data Science.

I. INTRODUCTION

If we look back in the history of science and technology, four scientific paradigms can be identified and distinguished [1]. Each new paradigm is a revolutionary improvement over the previous one, and it is based on invention and employment of entirely new and different set of tools and methods.

The first paradigm is the science based on empirical observations, and this is the era of Thales, Archimedes, Pythagoras, and Aristotle that continue with the invention of scientific methods which has been employed in the Middle Ages by Ibn al-Haytham and Roger Bacon.

The second paradigm of science and technology started in the mid-17th century with the invention of calculus, by Isaac Newton and Gottfried Leibniz, which forms the core of theoretical models and generalizations. This completely new tool makes a basis for scientific revolution and many new innovations and inventions including the results of James Clerk Maxwell and Albert Einstein.

Many scientific problems and the theoretical models became too complex with the time, so the analytical solution was no longer feasible [2]. With the advancement of computers in 1950, the third paradigm of computational science was born. This has allowed simulations of complex real-world phenomena based on the theoretical models of the second paradigm. Many products that we use today are developed using computer models and simulations like microprocessors, planes, rockets and many more.

The fast development of digital and computer systems in the last few decades, and especially with advancement in the data storage systems resulted in the creation and permanent storage of an enormous amount of data. This data stored in computer readable format recently has given rise to the fourth paradigm, which is known as data exploration or e-Science [1], and it unifies the first three paradigms of empirical observations, theory, and computation and simulation. The term e-Science is often defined as a coupling between accurate information tools, science, and scientists, whereas many of the practical scientific processes are guided and directed by the available data in each domain.

The data-driven science for the first time in the scientific history gives a chance to someone other than human to create new models and programs. With the progress in machine learning algorithms, now it is possible computers to learn the dependencies between parameters and to enable models' creation in areas that were impossible before. Some remarkable examples are models for personalized medicine, human speech recognition, playing Chess or Go, or even driving a car. This fast expansion of the machine learning is tightly connected to the availability of a considerable number of open source and free or commercial libraries and tools that simplify the process of using machine learning techniques [3]. Regarding the programming language popularity in the field of machine learning, Python is a leading one and is followed by R, Java, JavaScript, C and C++ [4].

Internet of Things (IoT) is another important trend that is happening in the last decade, and it is heavily contributing to a faster transition of science through its fourth paradigm. The IoT describes the connection of devices to the internet using embedded software and sensors to communicate, collect and exchange data with each other. IoT combines connectivity with sensors, devices, and people, enabling a form of a free-flowing conversation between human and machine, software and hardware [5]. Furthermore, forecasters predict there will be approximately 30 billion connected things by 2020 [5] These connected things include wearable sensors; smartphones; sensor-embedded gaming systems, such as music players; and in-vehicle sensor devices. Users carrying mobile sensing devices are becoming the source of

Dimitar Trajanov, Vladimir Zdraveski, Riste Stojanov and Ljupco Kocarev are with the Faculty of Computer Science and Engineering – ss. Cyril and Methodius University - Skopje, Ruger Boskovik 16 Skopje Macedonia, E-mail: {dimitar.trajanov, vladimir.zdraveski, riste.stojanov, ljupco.kocarev}@finki.ukim.mk

Ljupco Kocarev is also with Macedonian Academy of Science and Art, Bul. Krste Misirkov 2, P.O. Box 428, Skopje, Macedonia, E-mail: lkocarev@manu.edu.mk

a vast amount of varied data. At the same time, inherent user and device movement and ubiquitous connectivity are creating opportunities for dense spatial and temporal sensing.

Fig. 1. Multiple applications that create big data and percentage of data that will be transmitted and used (Source: Cisco Global Cloud Index, 2015–2020 [7])

This large number of IoT devices by 2020 will generate 600 ZB per year by 2020, 275 times higher than projected traffic going from data centers to end users/devices (2.2 ZB); 39 times higher than total projected data center traffic (15.3 ZB) [7]. Unfortunately, in today's IoT applications only less than one percent of the data is examined, and more than 99 percent of collected data is lost before reaching operational decision makers [8][7]. Most of the data that is actually used, for example, in manufacturing automation systems are used only for real-time control or anomaly detection. A great deal of additional value remains to be acquired, by using more of the data, linking this data to other data sets, as well as deploying more sophisticated IoT applications, such as using performance data for predictive maintenance or analyzing workflows to optimize operating efficiency. So, IoT can be a key source of big data that can be analyzed to capture value, and open data, which can be used by more than one entity [8]. This unused and hidden data that have a potential of creating new values is known as Dark Data.

Based on all previous findings and trends, the main goal of this paper is to raise awareness of opportunities and challenges that come from utilizing IoT created Dark Data for new applications that have a potential of scientific improvement of our lives or to increase productivities and effectiveness in the companies.

II. Dark Data in IoT: Concepts and Supporting Technologies

Gartner defines dark data as "the information assets organizations collect, process and store during regular business activities, but generally, fail to use for other purposes." It includes all data objects and types that have yet to be analyzed for any business or competitive intelligence or aid in business decision making [9]. Organizations often retain dark data for compliance purposes only.

Although this Dark Data may initially appear irrelevant, this information represents a large portion of available opportunities that many companies ignore. There is a substantial risk that comes with not addressing this data. If a company decides not to invest in the analysis and processing of dark data while its competitors do, the company will fall behind. Investing in dark data is an opportunity that greatly outweighs the cost [10].

A. Sources of data and data variety

Traditionally, most organizations primary collect transactional data used to run their business like orders, inventory tracking, customers interactions, operations or financial transaction. The internet services have created some entirely new sets of data that includes social networks data, pictures, videos, and a lot of textual information. But the real explosion of data variety is happening with the establishment and massive development of IoT [11]. The IoT devices are embedded into many objects that we are using today, with the expectation that they will be a part of every object in the near future.

In contrast with computers, the IoT devices are equipped with a broad spectrum of different sensors that gives the opportunity to sense many parameters of the physical environment. Some of the more common sensor types that are part of today's IoT are sensors for: Sound, audio, and acoustics; Pressure and force; Optic, light and imaging; Temperature and thermal; Motion and velocity; Flow, liquid, chemical and gas; Magnetic; Air, water and land pollution; and Proximity, position and presence. Because all those sensors are spatially distributed, the spatial information can be associated with the measurements.

The real-time data from sensors usually is used for alarms or real-time control. To be used in more applications, like optimization and prediction, data from sensors need to be stored at discrete time intervals and then used in new advanced applications. A typical industrial practice involves acquiring one measurement per second from each IoT device [12]. With this resolution of measurements, there will be 60 measurements per minute, 3,600 measurements per hour, leading to 31.5 million measurements per year. If we have an industrial facility with 3,000 parameters that are measured by IoT devices, then 94.6 billion measurements are generated per year. The net result is huge volumes of both structured and unstructured data that need to be analyzed to reveal patterns and associations.

Another type of data that is a real candidate for dark data are audio, video and image files. Contents of these media files, such as the people in the recording or dialogue within a video, will remain locked within the file itself if it is not processed.

Combining operational business data with sensor-generated information has a potential to deliver innovative ways to drive high performance and intelligent decision-making in every industry. So, the real power comes from

the data integrations and building more complex models of the real world.

B. Dark Data and IoT Intersecting Technologies

From the technological point of view, several other emerging technologies and trends are closely related to the Dark Data in IoT. The three concepts are worth to be mentioned here and have or will have a significant impact on improving the data usage in the IoT domain. These technologies are Industry 4.0, Digital Twins and Data Lakes.

In recent years, there have been great advances in Industrial Internet of Things (IIoT) and its related domains, such as industrial wireless networks (IWNs), big data, and cloud computing [14]. These emerging technologies are introducing advanced digitalization within factories, and in combination with future-oriented technologies in the field of "smart" objects (machines and products) have resulted in a new fundamental paradigm shift in industrial production. The future production will be based on modular and efficient manufacturing systems in which products control their own manufacturing process. That is supposed to accomplish the manufacturing of individual products in a batch size of one while maintaining the economic conditions of mass production. Motivated by this future expectation, the phrase "Industry 4.0" was established for a planned "4th industrial revolution" [15]. While in Industry 3.0 the focus was on the automation of single machines and processes, in Industry 4.0 the central goal is the digitization of all physical assets and integration into digital ecosystems with value chain partners.

One of the Gartner top 10 strategic technology trends for 2018, closely connected to the IoT and Dark Data, are Digital twins [16]. A digital twin is a digital representation of a real-world entity or system. The "twin" is the "digital" transformation of a real-world object or system which can be visualized and controlled by an analyst or manager in a manner that is location agnostic [17]. In the context of IoT, digital twins are linked to real-world objects and offer information on the state of the counterparts, respond to changes, improve operations and add value.

From a simulation point of view, the Digital Twin approach is the next wave in modeling, simulation and optimization technology. In the last decades, simulation become a state of the art technology, but it is restricted to a computer and numeric experts that use standard tools to answer specific design and engineering questions [18]. In this direction, digital twin opens two new challenges. The first one is a design of new algorithms and methods for real-time simulations of a variety of different physical objects and processes. The second challenge is how to utilize the collected data from digital twins to improve business and production processes in a company or our lives.

Digital twin in an IoT platform contains two categories of data. The metadata which does not change and includes the details that describe the device such as serial number, firmware version, device location, type of sensors attached, sensors precision, and some other parameters like model or year of manufacturing. The second category of data is real-time sensor and status data of the device.

If we look at the sources for dark data, we can note that it can be represented as any data types: structured, semi-structured or unstructured data. In order to store and manage this data, we need a new storage concept that will support all data types and additionally will support storage of streaming data that come from IoT devices. The methodology that supports all these requirements is the "Data Lake."

A "Data Lake" is a methodology enabled by a massive data repository, based on low-cost technologies, which improves the capture, refinement, archival, and exploration of raw data within an enterprise. A data lake contains raw unstructured or multi-structured data that for the most part has unrecognized value for the firm [19]. The data lakes are typically built to handle large and fast arriving volumes of unstructured data (in contrast to data warehouses' highly structured data) from which further insights are derived. Thus, the lakes use dynamic (not pre-build static like in data warehouses) analytical applications. The data in the lake becomes accessible as soon as it is created (again in contrast to data warehouses designed for slowly changing data). The data lake strategies can combine SQL and NoSQL database approaches and online analytic processing (OLAP) and online transaction processing (OLTP) capabilities [20]. The data lake has emerged as the recognized mechanism to enable organizations to define, manage and govern the use of various big data technologies. That represents an evolution of big data towards the mainstream use in an enterprise and the associated focus on management of such assets [24].

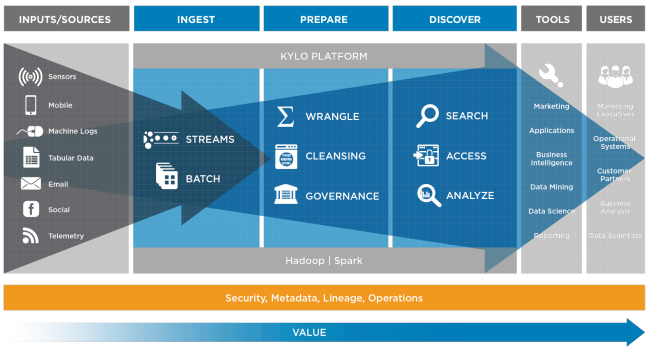

Fig. 2. Kylo Data Lake Architecture

To dynamically handle the structure of stored data, the Data lakes need to include metadata or semantic model of the data that adds a layer of context over the data defining the meaning and interrelationships with other data. Usually, the technologies of the Semantic web that are standardized by W3C are used, enabling easy integration of data stored

in the Data lake with Linked Data Cloud or other semantic web sources [21].

All major cloud providers like Microsoft [22] and Google [23] are offering a Data Lake storage. There is also high-quality open source enterprise-ready data lake management platform like Kylo[1], that enables self-service data ingest and data preparation with integrated metadata management, governance, security and best practices inspired by Think Big's 150+ big data implementation projects. The architecture of the Kylo platform is shown on Fig 2.

C. Security issues

Storing and securing data sometimes can gain more expense (and sometimes greater risk) than value. When both the location and the contents of files are unknown, they can be easily mined by hackers and used against a company. This can result in legal action, high payouts, and lost businesses [13]. Dark data can be a security risk and a barrier to operations when companies are unaware that the data even exists.

From the other side, IoT devices generate vast amounts of data on a daily basis that remain trapped in private infrastructures, due to the inability to control access to them and to connect this data with the rest of the world. By nature, this data is heterogeneous and sensitive, revealing the context, habits, and behavior of the owners. Hence, the publication of such data requires flexible security policies to control the access and interaction rights. However, in emergency scenarios, targeted disclosure of the private information is often needed in order to provide timely response [26].

Traditionally, the protection of the data generated by IoT devices is mainly based on securing the communication channel, most often through Transport Layer Security (TLS) or its Datagram TLS (DTLS) variant [28]. Additionally, the OAuth protocol is often used to limit the access to the available data and services, but it is not possible to include a context that plays a key role in this domain [29]. The IoT devices generate streams of data, and the work in [30] presents how these streams can be protected with centralized policies.

Even though there are models for different aspects of the IoT authorization, such as stream protection, context awareness, information flow control and identity providing (with certificates or OAuth), there is no complete solution that provides policies that cover all these features together. Moreover, the attacks that take over the IoT are becoming more often, introducing the need for device discoverability protection and configuration control. None of the analyzed solutions provides overcoming the heterogeneity in the IoT domain in the process of data protection. Among these challenges, a complete policy model should also cover all the features from the traditional enterprise (API based) systems, since the IoT devices are coordinated and consumed by this kind of applications, and the policies should provide distributed and complete protection of the whole infrastructure

III. Dark Data Collection and Usage Methodology

For any business, data is vital, because it holds the key to successfully manage the company, to attract new customers and increase growth. That is why the big data is big business. Dark data is not just a small portion of big data. It is the biggest slice of the pie and holds a massive amount of potential for those who can control it [13]. But, the central point to realize about dark data is that it does not have to stay dark. At the moment when dark data is used to gain insights, the data becomes actionable and is no longer dark.

D. How to start and build-up on current dark data.

In many cases, the organizations are just not aware of the dark data existence. So, in the beginning, there is a need to raise the awareness of existence and opportunities that can come from the dark data. Afterwards, the infrastructure that will support dark data analytics needs to be put in place. Creating a Data Lake infrastructure is the preferred solution, where gigabytes of data will be moved from multiple locations. This new storage will keep all data in one integrated system, where it will be easy to access and not to be forgotten again.

Based on our previous experience in many data-oriented projects [21][26][27][32][36] the following methodology is proposed:

1. *Get access*. Getting administrative access to everything, including all servers, hard drives and any other storage facilities used
2. *Search for data*. Search and identify all available data sources. Look at the applications, devices, peoples, and processes.
3. *Catalog data*. Analyze and categorize all data that is used by identified data sources, including the data stored in relational databases, logs, text data, multimedia data, IoT streams, IoT metadata, auditing data, and any other data that is stored.
4. *Security and privacy*. In this step, all legality issues need to be identified, and for all datasets, the assessment of security and privacy issues need to be conducted.
5. *Determine the value*. Based on the business needs determine which questions are the most important to be answered first. Identify datasets that will support answers to these questions.
6. *Move the data*. Store all or most of the data in the centralized Data Lake.

[1] https://kylo.io/

7. *Expand the data*. In this step, the goal is to find if there is additional important data that is sensed or collected but not stored. Examples include: some sensor data, intermediate data, additional more detailed log data, or data that is present but is not digitalized. These actions will require additional effort, so some estimate of the value of this data related to the price of getting it will be needed.
8. *Interlink the data*. Data that was collected come from different applications and sources and usually is not interlinked. We need to keep in mind that not only data but also relations carry information. In many cases, this information can be crucial for the business processes and models because it connects two or more different parts of the business.
9. *Link to external data*. Link the data with the external data sources like weather conditions, geolocations, stock exchange, news, large public and open data sets like DBpedia or Wikipedia.
10. *Create new data-driven applications*. Based on the business needs create new data-driven applications. In this process, usually, statistics and machine learning can be used to analyze the data (clustering, PCA, anomaly detection, novelty detection) or to create new models that will be used for predictions. Special emphasis needs to be given to data visualization in order to most effectively communicate the results with the users.

E. How to support future data based services

When you are designing a new service or application, there are two important issues associated with data that need to be followed. First, the collected data should be fine-grained as possible, meaning that you need to collect all possible details that are measured or are available. You can always get summaries and aggregate data, but there is no possibility to go back and derive more detailed data. The second principle is that data need to be available in real time. There is always a possibility to batch data or to slow down, but it cannot be speeded up.

The new developments in IoT like Industry 4.0 and Digital Twins are creating a lot of digital data. Only small portions of this data are available in factories or in our homes, so currently seems that for a big part of this data there are no ideas how to use it. But this does not mean that all these unusable details called “digital exhaust” do not need to be stored [11]. The idea behind the usage of Data Lake as a storage option is to allow an easy way to store many different types of data. This approach significantly reduces the effort that is required to store all available details. The Data Lake allows collection of data for future needs before it’s possible to know what those needs are, so it has tremendous potential. Data is not limited by the scope of thinking present when the data is captured but is free to answer questions we do not know how to ask yet: “Data itself is no longer restrained by initial schema decisions, and can be exploited more freely by the enterprise [24].

Having a lot of details about your business explores the opportunities to create new models and algorithms that would improve the business outcomes. In many cases, there is a general perception that models and algorithms are extremely complicated, but in many cases, there can be a very simple connection between business variables [11].

It is clear that there are real costs associated with the collection, transmission, processing, and storage of data. Some of the new technologies like Data Lake simplify the process and thus lower the costs, but still, there is a need for thoughtful rationalization of how much data is enough.

Processing a large data without a specific purpose in mind will possibly lead to failure. Indeed, dark analytics efforts that are surgically precise in both intent and scope often deliver the greatest value. Like every analytics journey, successful efforts begin with a series of specific questions. What problem are you solving? What would we do differently if we could solve that problem? Answering these questions makes it possible for dark analytics initiatives to illuminate specific insights that are relevant and valuable [25].

IV. DARK DATA UTILIZATION EXAMPLES

In the past couple of years, we have worked on several IoT connected projects where some forms of dark data were used. Three of our projects will be presented in more details, each of them covering different aspects of using dark data. (1) In the first project is a Data Lake platform for Smart City that was created to support storage and analytics of variety of different data types including industrial and personal IoT devices. (2) In the second presented project is about Power Grid Analytics and the concept of interlinking the data with external data sets are shown. (3) The last project shows the creation of metadata in the form of Ontologies and how analysis of dark data that is stored in semi-structured text files can be utilized to automate the process of System of chip synthesis.

A. Smart City Platform

We have developed a concept of a platform that complements the slow low-resolution (annual, quarterly or monthly) city's ISO indicators with high resolution, sensing, social media extracted knowledge and person-centric indicators, which can provide real-time input into models to infer or forecast future smart city indicators. The storage of the data is organized like Data Lake allowing variety of data types to be stored. Therefore, the software solution provides methods for observing and measuring phenomena of common interest (e.g. traffic conditions, air pollution, noise in urban areas), over large geographic areas, which exploits the inherent mobility of sensing devices. By combing with the data from social networks, news, blogs and other data sources this creates a generic

city's footprint that could guide solutions and technologies towards smart city's transition [31].

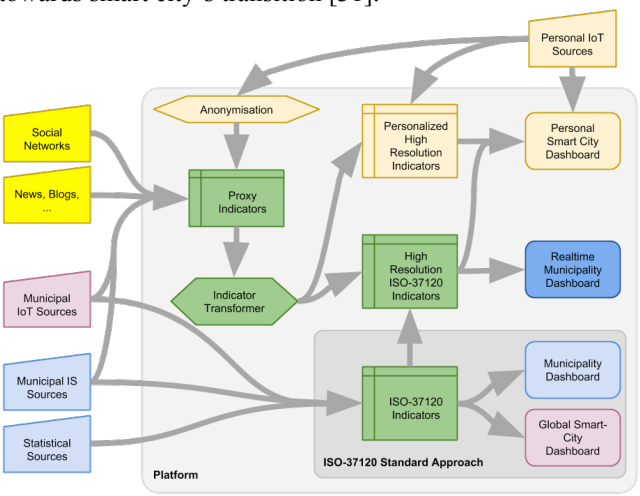

Fig. 3. Smart City Platform Architecture with Data Lake Storage

The core of the architecture, shown in Fig 3., is based on the ISO-37120 indicators (or ISO4City, ISO4C) and extended to a more granular (temporal, spatial) set of high resolution ISO-37120 indicators that provide transformation of the core indicators towards the real-time personalized dashboard. To tune the data, additional customization and contextualization to a viewpoint of a citizen is performed with the personalized high resolution indicators' filter, which is directly affected by the personal IoT sources. The data coming from users passes through anonymization filter, then the whole dataset is reduced through the proxy indicators filter and transformed towards the high-resolution ISO-37120 indicators, using the indicator transformer. The block Personal IoT sources represents integration with the personal (citizen's owned smart devices) IoT sensors. The specific Personal IoT block is designed to support the integration of personal IoT devices to the platform. The Personal IoT module architecture is component-based and enables easy development of new cartridges to support the integration of different classes of IoT devices. Integration of personal IoT raises the questions of privacy and security of the personal data.

B. Power grid analytics

The power grid ontology (PGO), shown in Fig 4, introduces a data model for power distribution system's data annotation. PGO is developed on top of schema.org and reuses and inherits many entities and properties. However, for the most specific domain requirements new entities and properties are introduced. The core entity is Node that represents a node in the power network, such as generator, substation, pillar of a transmission line and a power meter. Generators could be renewable such as wind turbine, solar, biomass, geothermal and hydro turbine and nonrenewable such as nuclear, coal, natural gas, crude oil and petroleum. After a sufficiently large data set is annotated with the PGO, it could be used to generate (periodically and continuously) a set of reports, such as to find the most over-loaded node or transmission line or to find the nodes with variable frequency or most frequent voltage drops. Several power grid data sets have been published recently, such as SciGRID[2] and GridKit[3] , that we have used to evaluate our concept and the ontology itself.

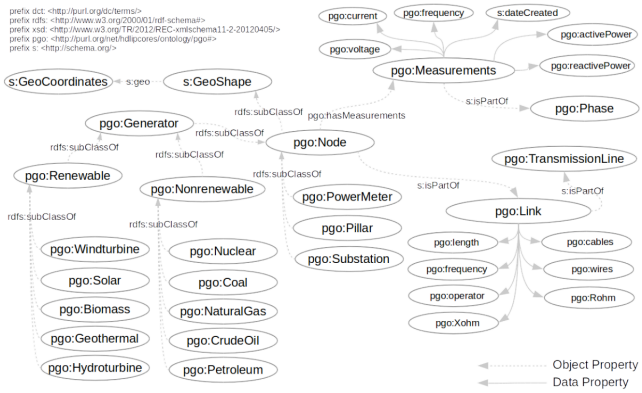

Fig. 4. Power grid ontology provides a data model for power distribution system's data (such as network topology and continuous and periodic measurements) annotation

As an example of linking data form the Power Grid with external data sources (DBPedia[4]), and introducing more granular indicators, we could find the for-example cities placed in radius of 20 km (or any other distance) from a power plant and calculate the total population living around the power plants. The results from this interlinking is shown on Table 1.

Table 1 - Cities in radius of 20 km from a power plants

Power Plant	Distance (m)	City	Population
Gemeinschaftskraftwerk Kiel	3021.53	Kiel	240832
Koepchenwerk	5607.79	Hagen	191241
Koepchenwerk	11579.8	Dortmund	575944
Koepchenwerk	18027.1	Bochum	361876
Kraftwerk Scholven	10144.6	Bottrop	117450
Kraftwerk Scholven	11287.4	Gelsenkirchen	260900
Kraftwerk Scholven	15564.4	Herne, N. R.-W.	166187
Kraftwerk Scholven	15747.0	Oberhausen	214990
Kraftwerk Scholven	16571.7	Essen	589075
Kraftwerk Scholven	19545.8	Bochum	361876
Statkraft Kraftwerk Knapsack II	11083.7	Cologne	1057327
TOTAL			4137698

C. HDL IP Cores system

In the HDL IP cores system first we have created meta data in the form of Ontologies, and we designed the system for synthesis of System on Chip that utilize the dark data that is extracted and analyzed from semi-structured HDL source text files.

[2] SciGrid, http://scigrid.de/
[3] GridKit, https://github.com/bdw/GridKit
[4] DBPedia SPARQL endpoint, http://live.dbpedia.org/sparql

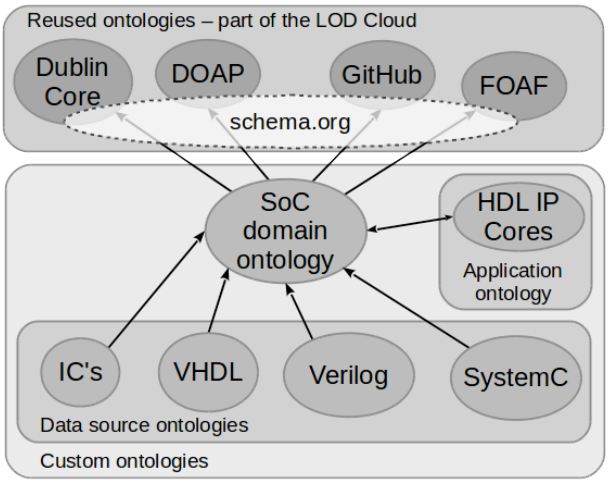

Fig. 5. SoC design ontologies

The Linked Data best practices [32] and the concept of semantic systems development, explored in [33], states that for any semantic system there must be a domain ontology, describing the knowledge in the given domain, a data-source ontology, strongly related to the type of the data source, and an application ontology, which describes the application entities and relations. Following this paradigm, we developed a set of ontologies, shown in Fig 5.

The ontologies on the top of the cloud in Fig 5. have already been mentioned and the ontology in the middle (SoC) is our domain ontology. It serves as a basic hardware architecture description schema, sufficient to annotate an existing system on a chip. At the bottom, there are ICs, VHDL, Verilog and SystemC data source ontologies, each extending the main hardware concepts from SoC, but providing additional language specific schema, suitable for its specific HDL. At the right-hand side of Fig. 5. SoC design ontologies, the HDL IP Cores application ontology (hipc.owl) is shown. It covers all classes and relations required to deploy a functional application and with that offers client-side features which provide a novel concept in the storage and retrieval of HDL IP Cores [34][35][36].

V. Conclusion

In this paper, the current trends and technologies related dark data in the IoT domain were presented. In order to utilize the benefits, of data driven science and machine learning, we need to get access and utilize all available data. To apply machine learning to your domain, you need a lot of data. There are two solutions for that. First is to create new applications that collect all data that is required to train machine learning based models, then to deploy the application, collect data for some time, and at the end use the collected data as an input to a machine learning algorithm. The second approach is to identify all white and dark data that is already in the system, to collect them and then to try to create machine learning algorithms based on the currently collected data. The advantage of the second approach is that it is cheaper, faster and what is more important it can provide access to a very large number of historical data.

Based on current trends in science and technology, companies first need to manage to store as much data as possible, and then they find a way to use this data creatively and innovatively that will give them a key completive advantage.

References

[1] T. Hey, S. Tansley, K. M. Tolle, et al., The fourth paradigm: data-intensive scientific discovery, vol. 1. Microsoft research Redmond, WA, 2009.

[2] Agrawal, Ankit, and Alok Choudhary. "Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science." APL Materials 4, no. 5 (2016): 053208.

[3] Joseph Misiti, A curated list of awesome Machine Learning frameworks, libraries and software, Available online at: https://github.com/josephmisiti/awesome-machine-learning, Accessed: 10 Jan 2018.

[4] Jean Francois Puget,"The Most Popular Language For Machine Learning Is", IBM Developer Works Blog, https://www.ibm.com/developerworks/community/blogs/jfp/entry/What_Language_Is_Best_For_Machine_Learning_And_Data_Science?lang=en. (2016)

[5] Ernst & Young Report, "Internet of Things: Human-machine interactions that unlock possibilities", Ernst & Young Global Limited, (2016).

[6] A. Nordrum, "Popular Internet of Things Forecast of 50 Billion Devices by 2020 Is Outdated," IEEE Spectrum, 18 Aug. 2016; http://spectrum.ieee.org/tech-talk/telecom/internet/popular-internet-of-things-forecastof-50-billion-devices-by-2020-is-outdated.

[7] Cisco White Paper, "Cisco Global Cloud Index: Forecast and Methodology, 2015–2020", Cisco Public, 2016

[8] Bughin, J., J. Manyika, J. Woetzel, M. Chui, P. Bisson, and R. Dobbs. "The Internet of Things: mapping the value beyond the hype.", McKinsey Global Institute (2015).

[9] Sony Shetty, How to Tackle Dark Data, Gartner, (2017), https://www.gartner.com/smarterwithgartner/how-to-tackle-dark-data/

[10] Frank Moreno, "Dark data discovery: Improve marketing insights to increase ROI", IBM Business analytics blog, Sep 20, 2016, https://www.ibm.com/blogs/business-analytics/dark-data-discovery-improve-marketing-insights-to-increase-roi/

[11] Rossman, John. "The Amazon Way on IoT: 10 Principles for Every Leader from the World's Leading Internet of Things Strategies", Clyde Hill Publishing, 2016

[12] Fuhr, Peter L., Marissa E. Morales Rodriguez, Sterling Rooke, and Penny Chen. "Convergence and

Commercial Momentum-Industrial Internet of Things Evolution." InTech 2017, no. 2 (2017).
[13] Stephen Mackey, "The Rise of Dark Data and How It Can Be Harnessed", KDnuggets: Opinions, Interviews, Reports, https://www.kdnuggets.com/2016/03/rise-dark-data-how-harnessed.html (2016)
[14] Wan, Jiafu, Shenglong Tang, Zhaogang Shu, Di Li, Shiyong Wang, Muhammad Imran, and Athanasios V. Vasilakos. "Software-defined industrial internet of things in the context of industry 4.0." IEEE Sensors Journal 16, no. 20 (2016): 7373-7380.
[15] Lasi, Heiner, Peter Fettke, Hans-Georg Kemper, Thomas Feld, and Michael Hoffmann. "Industry 4.0." Business & Information Systems Engineering 6, no. 4 (2014): 239-242.
[16] Kasey Panetta, "Gartner Top 10 Strategic Technology Trends for 2018", Gartner, (2017)
[17] Datta, Shoumen Palit Austin. "Emergence of Digital Twins-Is this the march of reason?." Journal of Innovation Management 5, no. 3 (2017): 14-33.
[18] Rosen, Roland, Georg von Wichert, George Lo, and Kurt D. Bettenhausen. "About the importance of autonomy and digital twins for the future of manufacturing." IFAC-PapersOnLine 48, no. 3 (2015): 567-572.
[19] Fang, Huang. "Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem." In Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2015 IEEE International Conference on, pp. 820-824. IEEE, 2015.
[20] Miloslavskaya, Natalia, and Alexander Tolstoy. "Big Data, Fast Data and Data Lake Concepts." Procedia Computer Science 88 (2016): 300-305.
[21] Trajanov, Dimitar, Riste Stojanov, Milos Jovanovik, Vladimir Zdraveski, Petar Ristoski, Marjan Georgiev, and Sonja Filiposka. "Semantic sky: a platform for cloud service integration based on semantic web technologies." In Proceedings of the 8th International Conference on Semantic Systems, pp. 109-116. ACM, 2012.
[22] Ramakrishnan, Raghu, Baskar Sridharan, John R. Douceur, Pavan Kasturi, Balaji Krishnamachari-Sampath, Karthick Krishnamoorthy, Peng Li et al. "Azure Data Lake Store: A Hyperscale Distributed File Service for Big Data Analytics." In Proceedings of the 2017 ACM International Conference on Management of Data, pp. 51-63. ACM, 2017.
[23] Halevy, Alon Y., Flip Korn, Natalya Fridman Noy, Christopher Olston, Neoklis Polyzotis, Sudip Roy, and Steven Euijong Whang. "Managing Google's data lake: an overview of the Goods system." IEEE Data Eng. Bull. 39, no. 3 (2016): 5-14.
[24] Amber Lee Dennis, "Data Lakes 101: An Overview", Dataversity, August 25 2016, Available online at: http://www.dataversity.net/data-lakes-101-overview/
[25] Tracie Kambies, Paul Roma, Nitin Mittal, Sandeep Kumar Sharma, "Dark analytics: Illuminating opportunities hidden within unstructured data", Deloitte Insights Tech Trends 2017, (2017),
[26] Stojanov, Riste, Sasho Gramatikov, Igor Mishkovski, and Dimitar Trajanov. "Linked Data Authorization platform." IEEE Access (2017).
[27] Jovanovik, Milos, and Dimitar Trajanov. "Consolidating drug data on a global scale using Linked Data." Journal of biomedical semantics 8, no. 1 (2017): 3.
[28] A. Niruntasukrat, C. Issariyapat, P. Pongpaibool, K. Meesublak, P. Aiumsupucgul, and A. Panya, "Authorization mechanism for mqtt-based internet of things," in Communications Workshops (ICC), 2016 IEEE International Conference on, pp. 290–295, IEEE, 2016
[29] Fremantle, Paul, and Benjamin Aziz. "OAuthing: privacy-enhancing federation for the internet of things." In Cloudification of the Internet of Things (CIoT), pp. 1-6. IEEE, 2016.
[30] B. Carminati, E. Ferrari, J. Cao, and K. L. Tan, "A framework to enforce access control over data streams," ACM Transactions on Information and System Security (TISSEC), vol. 13, no. 3, p. 28, 2010.
[31] Zdraveski, Vladimir, Kostadin Mishev, Dimitar Trajanov, and Ljupco Kocarev. "ISO-Standardized Smart City Platform Architecture and Dashboard." IEEE Pervasive Computing 16, no. 2 (2017): 35-43.
[32] Schmachtenberg, Max, Christian Bizer, and Heiko Paulheim. "Adoption of the linked data best practices in different topical domains." In International Semantic Web Conference, pp. 245-260. Springer, Cham, 2014.
[33] Hebeler, John, Matthew Fisher, Ryan Blace, and Andrew Perez-Lopez. Semantic web programming. John Wiley & Sons, 2011.
[34] V. Zdraveski, M. Jovanovik, R. Stojanov, and D. Trajanov, "Hdl ip cores search engine based on semantic web technologies," ICT Innovations 2010, Communications in Computer and Information Science, vol. 83, no. 2, pp. 306 – 315, 2011.
[35] V. Zdraveski, A. Dimitrovski, D. Trajanov. "HDL IP Cores System as an Online Testbench Provider". Small Systems Simulation Symposium, Volume: 5th, Nis, Serbia, February 2014.
[36] V. Zdraveski, M. Jovanovik, R. Stojanov, D. Trajanov. "HDL IP Cores Searh Engine Based on Semantic Web Technologies".