Enabling big data exploration in dynamic contexts

Ada Bagozi, Devis Bianchini, Valeria De Antonellis, Alessandro Marini, Davide Ragazzi

Dept. of Information Engineering University of Brescia
Via Branze, 38 - 25123 Brescia (Italy)

Abstract. According to the Industry 4.0 vision, big data management is among the new challenges for the factory of the future. While many approaches have been developed to investigate data analysis, data visualisation, data collection and management, the impact of big data exploration is still under-estimated. In this paper, we propose an approach for big data exploration in a dynamic context of interconnected systems, such as the Industry 4.0 domain. The approach relies on three main pillars: (i) a multi-dimensional model, that is suited for supporting the iterative and multi-step exploration of big data; (ii) novel data summarisation techniques, based on clustering; (iii) a model of relevance, aimed to focus the attention on relevant data only.

Keyword: big data exploration, data summarisation, data relevance, Industry 4.0, smart manufacturing

1 Introduction

Big data management challenges raised from the abundance of real time data in Cyber-Physical Systems (CPS), enabled by the widespread diffusion of IoT technologies [1]. Data is emerging as a new industrial asset, to implement advanced functions like state detection, health assessment, as well as *manufacturing servitization* [2]. In this context, many approaches have addressed issues related to data collection and storage, analysis and visualisation. Nevertheless, big data exploration issues are still under-estimated. These issues are very relevant with reference to the principle of "human-in-the-loop" for Industry 4.0 applications [3], where operators are in charge of taking decisions in unknown situations, based on their long-term experience. Operators must be supported in managing the high volume of collected data in order to identify relevant insights on which these decisions will be based.

In this paper, we discuss the ingredients to enable exploration of real time data in a dynamic context of interconnected systems, where large amounts of data must be incrementally collected, organized and analysed on-the-fly. Such

SEBD 2018, June 24-27, 2018, Castellaneta Marina, Italy.

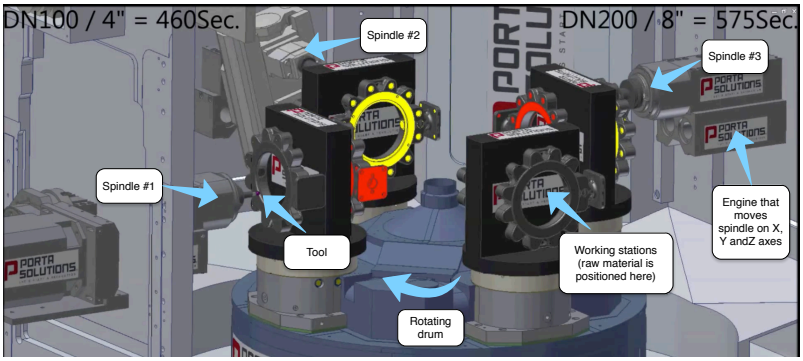


Fig. 1. The multi-spindle machine used for the case study.

ingredients have been already described in [10], where additional details about the proposed model and techniques, as well as an extended comparison with the state of the art, can be found. The approach relies on three main pillars: (i) a multi-dimensional model, to support the iterative and multi-step exploration of big data; (ii) novel data summarisation techniques, based on clustering; (iii) a model of relevance, aimed to focus the attention on relevant data only. The novel contribution of our approach relies on the clustering-based data relevance evaluation and its use in combination with the multi-dimensional model, in order to foster big data exploration. In fact, multi-dimensional data modeling, where information is organised according to dimensions, either flat or hierarchically organized, may ease data exploration [4]. Furthermore, data summarisation techniques enable aggregated views over high volume of data. Finally, data relevance evaluation helps operators to identify data of interest, also when exploration requirements are not well specified or must be iteratively refined based on collected data.

The paper is organised as follows: in Section 2 we will motivate the approach with the help of a smart factory case study and we highlight differences with respect to related work; in Section 3 we will provide a general overview of the approach; the architecture of a framework built on top of the proposed techniques and models will be presented in Section 4 and preliminary experiments in the case study will be described in Section 5; finally, Section 6 closes the paper.

2 Motivations and challenges

2.1 Smart factory case study

Let's consider an Original Equipment Manufacturer (OEM) producing multi-spindle machines. As shown in Figure 1, spindles work independently each other on the raw material, that is positioned on a rotating drum.
Each spindle is mounted on a unit moved by an electrical engine to perform X, Y and Z movements. The spindle rotation is impressed by an electrical engine

and its rotation speed is controlled by the machine control. Spindles use different tools (that are selected according to the instructions specified within the Part Program) in order to complete different steps in the manufacturing cycle. For each spindle, the velocity of the three axes (X, Y and Z), the electrical current absorbed by each engine, the value of rpm for the spindle and the percentage of power absorbed by the spindle engine (charge coefficient) are measured. Hereafter, with term *features* we will refer to the measured quantities. The aim is to monitor the axle hardening of each spindle and the tool wear. Axle hardening is monitored by observing changes in the values of energy consumption (spindle engine charge coefficient) for similar rpm. By detecting differences in energy consumption while using different tools, spindle hardening can be identified as the possible anomaly during the manufacturing operations. If the increase in energy consumption is related only to the usage of a specific tool, this has been recognised as a symptom of tool wear. We collected real data from three machines, each one equipped with three spindles and different tools. On each spindle, we monitored the eight features listed above.

2.2 Related work

In the considered motivating scenario, experience of human operators still plays a fundamental role. Data exploration must be performed in near real time and during incremental data collection. In [5] Exploratory Computing is defined as a multi-step process, going beyond traditional exploratory data analysis and Data Mining techniques. In our approach, we aim at performing a step forward by describing a combined use of multi-dimensional model, data summarisation and relevance evaluation techniques to foster exploration while dealing with disruptive characteristics of big data (e.g., mainly volume and velocity). The innovation we introduced, compared to On Line Analytical Processing [6] and faceted search [7], relies on the combined use of data summarisation and relevance evaluation techniques within a multi-dimensional model specifically designed for incremental big data organisation. In [4] OLAP-based exploration for multi-dimensional data is discussed, but no data relevance techniques are proposed. Similarly, authors in [8] propose the application of query approximation techniques for data that are incrementally collected, without providing a way of guiding target users towards data of interest. In [9] an approach to issue range queries over structured data is described. Semantic windows group data according to specific criteria (e.g., all data in a given time interval, all data in the same geographical area), while sampling techniques are applied to support users in query resolution. No incremental collection and organisation of data are performed. The three pillars we will introduce in the following (clustering, data relevance and multi-dimensional modeling) have been combined to foster exploration of huge amounts of data incrementally collected from dynamic systems.

3 Approach overview

Data summarisation. The characteristics of big data, namely, volume, velocity and variety, pose non trivial issues for data collection and organisation. High

volume calls for techniques and tools to provide a compact view over the large amount of collected data. Furthermore, we deal with data streams, in which data is collected in a fast and incremental way. We address these issues by applying incremental clustering techniques. Clusters offer a two-fold advantage: (a) they give an overall view over a set of measures, using a reduced amount of information; (b) they allow to depict the behaviour of the system better than single records, that might be affected by noise and false outliers, in order to observe a given physical phenomenon. Clustering is performed in two steps: (i) in the first one, a variant of Clustream algorithm [11] is applied, that incrementally processes incoming data to obtain a *set of syntheses*, that provide a lossless representation of measures; (ii) in the second step, syntheses are clustered, in order to minimise the distance between their centroids within the same cluster and to maximise the distance between centroids across different clusters. Clusters give a balanced view of the observed physical phenomenon, grouping together syntheses corresponding to the same working status. The two-steps clustering algorithm enables an incremental procedure specifically developed to face data streams. Considering Δt as the time interval in which measures are grouped in syntheses, that in turn are clustered, every Δt seconds a new clusters set SC is generated, built on top of the previous iterations.

Clustering-based data relevance evaluation. Relevance-based techniques are used to focus exploration on relevant clusters only. In literature, data relevance is defined as the distance from an expected status. In our case, the expected status corresponds to the normal working conditions of monitored Cyber Physical Systems. The expected status can be tagged by human operators while observing the monitored system. Let's denote with $\hat{SC} = \{\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_n\}$ the clusters set identified during normal working conditions, and with $SC = \{C_1, C_2, \ldots, C_m\}$ the current clusters set, where n and m do not necessarily coincide. Relevant data are recognised when SC differs from $\hat{SC}$. Therefore, the proposed relevance techniques are based on clusters set distance between SC and $\hat{SC}$, denoted with $\Delta(SC, \hat{SC})$, and enable to detect clusters movements, clusters contraction/expansion, changes in the number of clusters. We refer to [10] for details on $\Delta(SC, \hat{SC})$ computation.

Multi-dimensional model for data exploration. The multi-dimensional model is shown in Figure 2, using the hypercube representation, where axes represent exploration dimensions and nodes represent clusters sets computed on measures. Multi-dimensional modeling may help to organise clustered data according to different perspectives, compliant with operators' requirements. Among the dimensions, we always consider feature spaces, time and operational parameters. With *feature spaces* we refer to monitored phenomena (e.g., axle hardening, tool wear) observed by measuring a set of features. Time is of paramount importance, since the data summarisation procedure and data relevance techniques are strictly related to the time dimension. Operational parameters represent domain-specific settings or variables: for example, the working mode (G0, fast movement

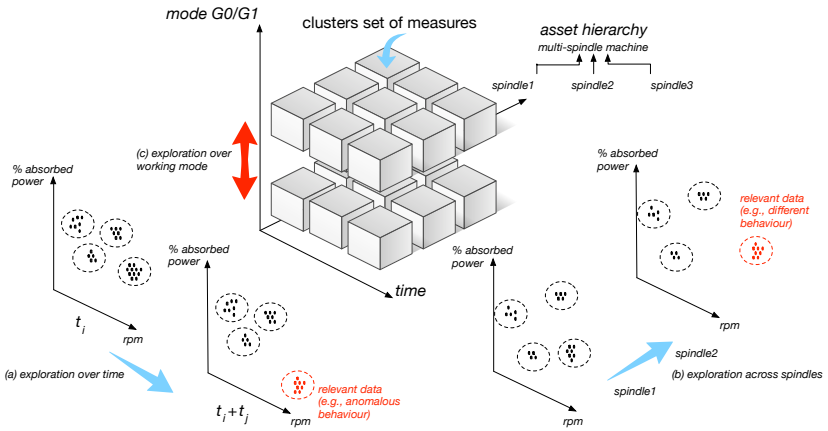


Fig. 2. The multi-dimensional data model for big data exploration.

of the spindle to catch the tool, or G1, slow movement of the spindle during the manufacturing), the tool used during manufacturing, the monitored system, the manufactured product. Dimensions may present hierarchies: for example, tools can be aggregated into tool types, monitored system follows the asset hierarchy (enterprise, shop floor, machine, components).

Exploration scenarios. Data relevance is used to identify dimensions over which summarised data changed, thus denoting an unexpected behaviour of the monitored system. Nevertheless, data relevance evaluation performed over all available dimensions may bring to high computational requirements. Therefore, *exploration scenarios* are proposed to constrain relevance evaluation over a subset of available dimensions, to prevent useless comparisons. In Figure 2, some examples of exploration scenarios are highlighted. Exploration for anomaly detection (a) detects anomalies by observing if collected data of a specific spindle changes and gets closer/overtakes physical limits of breakage. This exploration scenario constrains exploration over the time dimension and within the same monitored system. Exploration for performance comparison (b) focuses comparison between spindles working in the same conditions. Other exploration scenarios are defined to avoid erroneous comparisons. For example, let's consider the exploration over working mode (c) as shown in Figure 2: a comparison of system behaviour across G0 and G1 modes does not make sense. Modeling exploration scenarios, in order to abstract from their implementation, will be further investigated in future work.

4 IDEAaS Architecture

The approach described in this paper has been implemented within the IDEAaS (Interactive Data Exploration As a Service) framework, whose architecture is

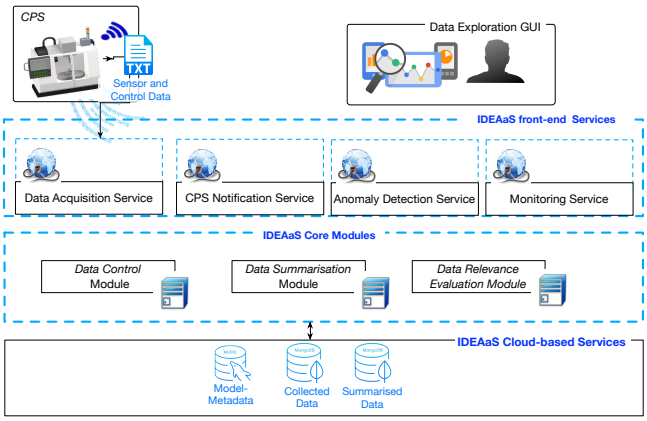
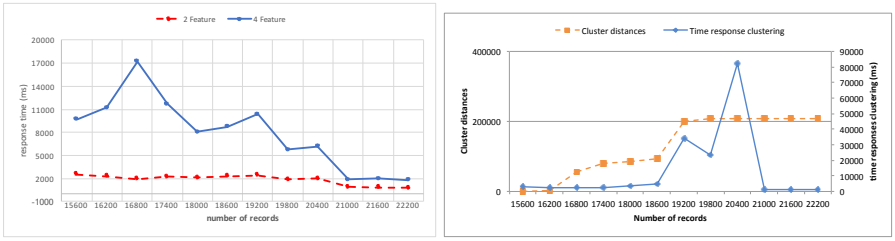


Fig. 3. The functional architecture of the IDEAaS framework.

shown in Figure 3. In the IDEAaS modular architecture, modules are distinguished in IDEAaS Core Modules (that include Data Control, Data Summarisation and Data Relevance Evaluation) and IDEAaS front-end Services, exposed both as web services and as standalone modules. Among these services, Data Acquisition Service is in charge of storing data collected from CPS. As shown in Figure 3, data coming from the physical system, properly collected through sensors and IoT technologies, is sent to the Data Acquisition Service to be processed. This service operates in order to minimise time spent for data acquisition. Specifically, data collected is first saved as JSON documents in a MongoDB NoSQL database (*Collected Data*), minimising any other operation, namely data control and cleaning and data clustering that are performed in parallel. Dimensions of the multi-dimensional model are stored as metadata into the *Model-MetaData* relational database. On collected data, summarisation and relevance evaluation techniques are applied. Resulting clusters sets are stored as JSON documents within *Summarised Data* database (also using MongoDB technology). The CPS Notification Service is in charge of notifying to target operators data of interest according to the relevance evaluation techniques. Data relevance can be used to serve different purposes, ranging from anomaly detection to performance monitoring of CPS, for which specific services will be designed in the future. On top of CPS Notification Service, a Data Exploration GUI will be developed as well, to allow operators to interact with the system and explore collected data. The IDEAaS architecture is implemented in Java, on top of a Glassfish Server Open Source Edition 4.

5 Preliminary experiments

We tested the efficiency and effectiveness of our approach in providing summarised data for exploration purposes, given the acquisition rate of records in

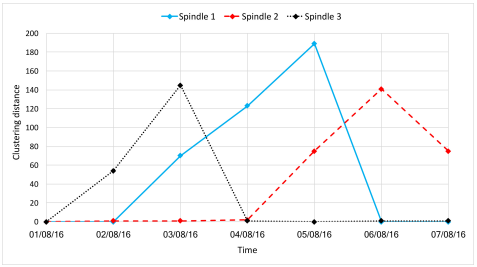


(a) (b)

Fig. 4. Tests on efficiency of clustering and hypercube generation (a) and on the effectiveness of the model of relevance, introducing a variation in collected records (b).

Working Days	Spindle 1		Spindle 2		Spindle 3	
	rpm decreas	energy consumption increase	rpm decreas	energy consumption increase	rpm decreas	energy consumption increase
01/08/2016 (operates normally)	/	/	/	/	/	/
02/08/2016	/	/	/	/	15%	16%
03/08/2016	20%	20%	/	/	29%	29%
04/08/2016	30%	30%	/	/	/	/
05/08/2016	40%	40%	20%	20%	/	/
06/08/2016	/	/	30%	30%	/	/
07/08/2016	/	/	20%	20%	/	/

(a)



(b)

Fig. 5. Introduced variation in collected records that simulates spindle hardening (a) and evidences of the relevance evaluation techniques to identify changed records (b).

the considered case study. We collected 140 millions of records from the three machines. All records present a timestamp and have been collected every 200ms (5 records per second). We run experiments on an Intel Core i7-6700HQ, CPU 2.60 GHz, 4 cores, 8 logical cores, RAM 16GB. Collected records of measures have been saved within MongoDB as JSON documents grouped into collections. Documents present a simple structure, with at most one level of depth, and collections have been organised considering the time as main dimension, in order to speed up both data storage and data extraction for clustering, that is applied on slots of records on a time interval Δt. This enabled to storage all 140 millions of records in 1 hour and 14 minutes, with an acquisition rate of ∼31,531 records per second. We tested clustering and hypercube generation on real data considering average values on 2 and 3 features (Figure 4a). The worse response time corresponds to the case where we performed clustering and relevance evaluation when no previous syntheses have been generated. Also in that case, IDEAaS framework was able to process ∼15,600 records in 11.5 seconds (processing rate of ∼1,356 records per second). Through the tasks of syntheses generation and clustering, the processed set of records is reduced to 7,2% on average. As shown in Figure 4b, we observed a variation in distance between clusters sets at the cost of decreasing the processing time to ∼255 records per

second, that is still acceptable. To test effectiveness of the IDEAaS framework in detecting working anomalies, we introduced unexpected working states to simulate spindle hardering by increasing energy consumption and decreasing rpm on a subset of collected data, as shown in Figure 5a. Experimental results depicted in Figure 5b show the clustering distance between clusters sets calculated at the time t and those calculated in case of normal working conditions. The figure shows how the techniques of the IDEAaS framework allow to timely identify the unexpected situations induced in the system under observation.

6 Concluding remarks

In this paper, we discussed the ingredients to enable exploration of real time data incrementally collected, organized and analysed on-the-fly. Our approach combines: (i) a multi-dimensional model, that is suited for supporting the iterative and multi-step nature of data exploration; (ii) efficient data summarisation techniques, based on clustering; (iii) a model of relevance, to focus the attention on relevant data only. Future efforts will be devoted to the parallelisation of clustering and relevance evaluation by relying on data organisation in the multi-dimensional model, to the development of data visualisation techniques and to the definition of different data exploration scenarios.

References

1. L. Monostori, Cyber-physical production systems: Roots, expectations and R&D challenges, in: Proc. of the 47th Conf. on Manufacturing Systems, 2014, pp. 9–13.
2. J. Lee, B. Bagheri, H. Kao, A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems, Manufacturing Letters 3 (2015) 18–23.
3. F. Longo, L. Nicoletti, A. Padovano, Smart operators in industry 4.0: A human-centered approach to enhance operators capabilities and competencies within the new start factory context, Computers & Industrial Engineering 113 (2017) 144–159.
4. N. Kamat, P. Jayachandran, K. Tunga, A. Nandi, Distributed and Interactive Cube Exploration, in: Proc. of 30th Int. Conf. on Data Engineering (ICDE 2014), 2014.
5. M. Buoncristiano, G. Mecca, E. Quintarelli, Roveri, D. Santoro, L. Tanca, Database Challenges for Exploratory Computing, SIGMOD Record 44 (2) (2015) 17–22.
6. M. Golfarelli, S. Rizzi, Data Warehouse Design: Modern Principles and Methodologies, McGraw-Hill, 2009.
7. D. Tunkelang, Faceted Search (Synthesis Lectures on Information Concepts, Retrieval and Services), Morgan and Claypool Publishers, 2009.
8. A. Wasay, M. Athanassoulis, S. Idreos, Queriosity: Automated Data Exploration, in: Proc. of the IEEE International Congress on Big Data, 2015.
9. A. Kalinin, U. Cetintemel, S. Zdonik, Interactive data exploration using semantic windows, in: Proc. of the ACM SIGMOD 2014, pp. 505–516.
10. A. Bagozi, D. Bianchini, V. De Antonellis, A. Marini, D. Ragazzi, Summarisation and Relevance Evaluation Techniques for Big Data Exploration: the Smart Factory case study, in: Proc. of 29th Int. Conference on Advanced Information Systems Engineering (CAISE 2017), 2017, pp. 264–279.
11. C. Aggarwal, J. Han, J. Wang, P. Yu, A framework for clustering evolving data streams, in: Proc. of 29th Int. Conf. on Very Large Data Bases, 2003, pp. 81–92.