FROM DATA WAREHOUSES TO ANALYTICAL ATOMS – THE INTERNET OF THINGS AS A CENTRIFUGAL FORCE IN BUSINESS INTELLIGENCE AND ANALYTICS

Henning Baars
Universität Stuttgart, baars@wi.uni-stuttgart.de

Julian Ereth
University of Stuttgart, ereth@wi.uni-stuttgart.de

Follow this and additional works at: http://aisel.aisnet.org/ecis2016_rp

Recommended Citation
Baars, Henning and Ereth, Julian, "FROM DATA WAREHOUSES TO ANALYTICAL ATOMS – THE INTERNET OF THINGS AS A CENTRIFUGAL FORCE IN BUSINESS INTELLIGENCE AND ANALYTICS" (2016). Research Papers. 3.
http://aisel.aisnet.org/ecis2016_rp/3
FROM DATA WAREHOUSES TO ANALYTICAL ATOMS –
THE INTERNET OF THINGS AS A CENTRIFUGAL FORCE
IN BUSINESS INTELLIGENCE AND ANALYTICS

Research

Baars, Henning, University of Stuttgart, Stuttgart, Germany, baars@wi.uni-stuttgart.de
Ereth, Julian, University of Stuttgart, Stuttgart, Germany, ereth@wi.uni-stuttgart.de

Abstract

For decades, the Data Warehouse (DW) has survived as a central architecture component in Business Intelligence and Analytics (BIA) landscapes despite a number of alternative concepts and continuous forces towards more decentralized structures. However, distributed concepts recently seem to have grown in relevance. This paper investigates causes and consequences of an increasing decentralization in order to derive recommendations for future BIA architectures. We have conducted a literature review for identifying relevant application drivers, challenges and building blocks for BIA solutions. The results suggest that particularly Internet of Things (IoT) applications drive federated analytical “ecosystem” solutions. In two case studies from the realm of Advanced Manufacturing we deepen the respective insights. On the one hand, our results underscore the business potential of integrated BIA solutions for IoT applications. On the other hand, the applications introduce a set of particular challenges that can be mapped to DW characteristics, mainly regarding data history, integration, administration, and governance. Based on the results, we project the identified trends into the future and suggest a federated, platform-based solution with “Analytical Atoms”.

Keywords: Business Intelligence and Analytics, Data Warehousing, Internet of Things, Advanced Manufacturing.
1 Introduction

The term “Business Intelligence and Analytics” (BIA) subsumes approaches to IT-based management and decision support (Chen et al. 2012). A core property of BIA is the striving for integration which in particular goes along with the provision of a Single Point of Truth (SPOT). A SPOT is supposed to ensure consistent and efficient decision making. In most BIA architectures, the SPOT is realized with a Data Warehouse (DW), an integrated, non-volatile, time-variant, and subject-oriented data repository (Inmon 2005). The application-independent nature of the DW gives it infrastructural characteristics in the sense of Keen (1993). Application-dependent data repositories for BIA are usually called data marts and can be realized as excerpts of the DW. Originally conceived to be used for managerial uses only, the DW has been extended to support operational decision making as well (Golfarelli et al. 2004). This necessitates near-real-time access to high-granular data – usually kept in an Operational Data Store (ODS) (Akbay 2006; Kemper and Baars 2009; Inmon et al. 1999).

![Figure 1: Exemplary DW-based BI architectures – textbook vs. reality](image)

Figure 1 illustrates the text-book layout of a DW-based architecture (Inmon 2005; Moss and Atre 2003; Chaudhuri et al. 2011) and an exemplary sketch of an actual architecture how it can be found in many companies. The deviations from the ideal have manifold causes like a) the inflexibility and the inherent complexity of DW architectures despite increasing agility pressures (Krawatzek and Dinter 2015; Li 2006), b) operations across enterprise borders (Baars and Kemper 2011), c) highly specific and latency-intolerant business requirements (Koch et al. 2010; Hänel and Felden 2011), and d) new approaches to store and analyze data under the label “Big Data” (Chen et al. 2011; Baars et al. 2014; Russom 2011; Li 2006). All these factors materialize in an increasing decentralization of BIA architectures with the risk of unwanted parallel DWs, unsupervised data marts, spreadsheet-based reporting, and/or an uncontrolled raw data access (Zimmer et al. 2012).

As a consequence of this, the suggestion of stripping the “classical” DW from the BIA landscape has been proposed – repeatedly. Usually this goes along with the idea to replace the DW with a direct access to raw data. Respective approaches have been brought forward under varying titles: Virtual DWs (Bernstein and Haas 2008), In-Memory databases (Plattner 2009), or recently “data lakes” (Terrizzano et al. 2015). However, so far the DW has weathered all those concepts. There are good reasons for this. To name a few: A decoupling from operational systems limits performance impacts, the data in a DW can be pre-processed for data-quality, access performance, and inter-source consistency. It can also be centrally governed by the help of administrative tools and for metadata management. Above all, the data in the DW can be versioned and be put in its historical context, by bringing in the time dimension, which is crucial for most analytical uses. In contrast, in operational
systems, both indicator and master data is repeatedly overwritten (Inmon 2005; Kemper 2000; Kemper and Baars 2009).

This all leads to a fundamental contradiction between the need for a reproducible and thorough pre-processing of decision support data (Chaudhuri and Dayal 1997) and the requirement of providing it in a versatile ad-hoc fashion with minimal latency in a decentralized BIA architecture. It is not yet clear in which domains this chasm will manifest, how wide it will become, and how it can be bridged. Hence, our studies address the three implied research questions:

1) What are relevant application areas in which decentralization gains in relevance?
2) What are the arising challenges that have to be addressed?
3) What are building blocks (architecture patterns and technologies) of solutions?

By searching for answers to these questions, we aim at deriving an architecture vision that helps to bridge this chasm. The conceptual framework in Figure 2 illustrates the relations of the before mentioned aspects and their arrangement in the context of the research.

In order to contribute to answers, we put our research into the context of the related work (section 2) and present a literature review on distributed decision and management support with respect to the research questions (section 3). The results are mirrored against two case studies which provide further insights and enrichments (section 4). On this foundation, we delineate the arising challenges (section 5) and sketch our architecture vision of federated “Analytical Atoms” (section 6). The paper concludes with a reflection of the limitations and the contribution of our research.

2 Related Work

Before further delving into distributed scenarios, it is reasonable to start with the defining requirements of a DW. These go back to Inmon (2005): non-volatility, subject-orientation, time-variance, and integration. Any future DW architecture would need to incorporate respective functionalities. These four characteristics can be further broken down by looking at the functionality of commercial DW software tools and at the data transformations necessary to convert data into information.

The particular DW variants in discussion are usually organizationally decentralized. This usually implies some level of distribution which refers either to the logical (distributed among independent processes) and/or the geographical (distributed across multiple locations) aspect of the architecture (Sheth and Larson 1990). Distributed structures that encompass various subsystems and data sources and possibly span across enterprise-borders are also often referred to as digital ecosystems (Chen et al. 2011; Liggesmeyer et al. 2014). Here, the cross-company trait is particularly noteworthy, as it introduces a variety of new challenges to BIA.
A general architectural pattern for handling distributed systems is data federation (White 2005) where multiple data sources are transparently mapped in order to appear like a single logical component (Heimbigner and McLeod 1985). For the realm of BIA and data warehousing, this leads to federated DWs (Berger and Schrefl 2008; Akinde et al. 2002; Inmon 2010). Unlike “classical” architectures, where data is consolidated within a single logical component (data consolidation), the decision-relevant data in a federated DW is kept in the source systems that are accessed directly and integrated on a logical level. The idea to do analysis and reporting directly on raw source data has also been brought forward under the labels “Enterprise Information Integration” (EII) and “Virtual Data Warehouses” (VDW) (Bernstein and Haas 2008).

The concepts for VDW/EII have entered commercial BIA tools for some time now, where they are particularly used for real-time scenarios with limited volumes of data. The same idea is also applied in In-Memory databases (Plattner 2009; Chaudhuri et al. 2011) that help to overcome performance restrictions that severely hampered older EII/VDW approaches before. In-Memory approaches can lead to significant performance gains, by removing performance-oriented data modelling needs and DW layers, and thus support a more user-driven, agile BIA (Knabke and Olbrich 2011). However, they also have shortcomings when it comes to the availability of historical data from external sources. Besides, it is often necessary to store the intermediate results of data transformations in the DW in order to comply with legal or contractual requirements, and to trace back a decision to its data source (data lineage / data heritage) (Winsemann and Köppen 2011).

Another architectural pattern for the emerging BIA ecosystems are platforms that abstract data sources and that handle the exchange, the storage, and possibly also the transformation of data. Contributions to this idea first came from Service Oriented Architectures (SOA) (Pospiech and Felden 2013; Dinter and Stroh 2009). White (2005) has coined this data propagation for the underlying principle. A similar approach is an enterprise-wide master data management (MDM) that also moves overarching functionality to a superordinate platform. In MDM, the definition, administration, and maintenance of master data is handled in an integrated fashion across applications (Loshin 2010; Otto and Reichert 2010). The relevance of a platform approach has been further corroborated by the emergence of Cloud-BIA services, i.e. Internet-based BIA services that can be provisioned ad-hoc and in a self-serving fashion. While still met by skepticism mainly because of concerns regarding data security, Cloud-BIA promises unprecedented degrees of scalability and agility (Thomson and van der Walt 2010; Ouif and Nasr 2011) and is particularly interesting for cross-company BIA solutions (Baars and Kemper 2011).

In the course of the ongoing Big Data discussion, new approaches like data lakes were brought forward. The data lake concept builds up onto the idea to indiscriminately collect raw data together with relevant metadata and store it in a separate Big Data Store like Hadoop. This idea leads back to a more centralized approach and resembles a DW. Nevertheless, there are some serious caveats that need to be considered when following a data lake approach, e.g. with respect to data quality and data collection (Terrizzano et al. 2015). Thus, Big Data solutions are currently not seen as a replacement for the existing DW but as an addition. This is also the way most BIA vendors and consultants include them in their current architecture blueprints. Apart from that, Big Data introduces a plethora of additional application-specific components, notably graph and NoSQL databases, which leads to a steep increase in the number of databases and technologies in the BIA landscape. Hence, instead of mitigating it, Big Data further impedes the integrated nature of BIA.

In conclusion, there are indeed a few relevant architectural patterns that have been proposed for distributed BIA ecosystems. However, it remains unclear where they come into play in the light of the ongoing developments on the application side (emerging application areas) and how these building blocks can be combined to a sustainable architectural vision that is grounded in relevant challenges on the application side.
3 Literature Review

The aim of the literature review was to systematically draw together knowledge regarding the subject in discussion. It particularly focuses on identifying application areas for federated, ecosystem-oriented, and possibly cross-company BIA applications, the challenges they introduce, and the relevant building blocks (technologies and architectural patterns) in order to connect them to an architectural vision.

On the method side, we followed Brocke et al. (2009) and Webster and Watson (2002). We queried four databases for academic publications relevant in the domain of Information Systems, namely: IEEE Xplore, the ACM Digital Library, SpringerLINK, and ScienceDirect.

<table>
<thead>
<tr>
<th>Federated</th>
<th>Ecosystem</th>
<th>Cross-company</th>
</tr>
</thead>
<tbody>
<tr>
<td>83</td>
<td>109</td>
<td>6</td>
</tr>
<tr>
<td>Analytics</td>
<td>90</td>
<td>84</td>
</tr>
</tbody>
</table>

Table 1. Number and the temporal distribution of reviewed articles

On the basis of an initial exploration of related work, we chose the keywords “Business Intelligence” and “Analytics” in each possible combination with the terms “federated”, “ecosystem” and “cross-company”. We chose those terms because, firstly, they appear to be overarching concepts and not temporary or technology-dependent buzzwords (Webster and Watson 2002) and, secondly, we deem them as being the prime novel aspects on the application side that we want to address. The databases were queried with a full-text search. The initial search results encompassed 1,652 academic articles published between January 2005 and September 2015. This includes the results from an additional backward search (Webster and Watson 2002). We removed duplicates and documents that we considered irrelevant, esp. when they did not fit into our conceptual framework, e.g. dealing with low-level implementation details. After that, 373 publications remained (cf. Table 1).

Table 2. Results of the keyword analysis

All results were manually screened to identify their primary subject and addressed concepts. Each article was assigned up to three main terms based on suggested keywords by the author(s) and the assessment of the reviewers. The terms were afterwards first iteratively condensed by merging synonyms or similar meanings (e.g. “ubiquitous” and “pervasive” computing) and second sorted...
according to our framework of reference. Table 2 depicts the results in a quantitative fashion. The temporal distribution reveals a constantly growing number of articles since 2012 and thereby emphasizes the perceived relevance of the subject.

As for the general theme in discussion, 41 of the 113 identified articles explicitly deal with some kind of “digital ecosystem” as introduced in Chapter 2. All these approaches share the idea of a network of “largely and loosely coupled actors that sense, respond, and adapt to each other through a common language, technology, and institutions to engage in an exchange of some sort” (Zeng and Lusch 2013). However, a closer look reveals that the subsumed concepts address different layers of abstraction. Some articles focus on the business or service side and understand a digital ecosystem as a collection of companies that interact and collaborate in order to create, manage, and offer new products and services (Hirsch et al. 2013; Hai et al. 2007). Others come from the logical architecture side and see digital ecosystems as a conglomerate of databases, sensors and other systems (Liggesmeyer et al. 2014; Zeng and Lusch 2013). A third, more solution-specific approach defines a digital ecosystem as a network based on “autonomous proactive (software) agents that act for their own advantage” (Barker and Finnie 2009). It needs to be acknowledged that the approaches are mutually complementing each other by bringing in different aspects of the ecosystem idea.

3.1 Emerging application areas

The most common application areas are from the domains of manufacturing (14 articles) and healthcare (14 articles). Within manufacturing, the main subject in discussion is the advancing digitization of factories and the associated impacts on infrastructures (Nahavandi et al. 2015; Nguyen and Savio 2008) as well as on future service models (Xinghao et al. 2014). In healthcare, many articles refer to digital healthcare ecosystems (Qureshi 2014; Serbanati et al. 2011) that integrate various systems and enable an information sharing between health institutions (Vargheese and Prabhudesai 2014) as well as ubiquitous healthcare over biosensors and mobile devices (Ravikumar et al. 2015). Further application domains that are addressed in the identified literature are Smart Cities, Smart Grids and other “Smart” Scenarios like the Smart House or Smart Cars. Altogether, 28 articles explicitly discuss “Smart” concepts. Though these are often defined in a rather vague fashion, they mostly incorporate the goal to use embedded digital and internet technologies to enhance quality and performance and/or to reduce costs and resource consumption (Giffinger et al. 2007). Respective solutions are also commonly subsumed under the term Internet of Things (IoT) that refers to the vision of IT-based networks of (smart) physical objects (Atzori et al. 2010). The application of IoT technologies to the non-consumer domain is discussed under the heading Industrial Internet (Industrial Internet Consortium 2015). The identified articles that deal with such IoT/Industrial Internet applications mostly discuss the necessary technological infrastructure (Maciel et al. 2015; Khan et al. 2012), issues of interoperability, and/or the general role of data and information in smart environments (Khan et al. 2015; Poncela et al. 2014; Carlton et al. 2015). Moreover, some articles also discuss analytical applications for managing supply chains, e.g. approaches for cross-company integration (Wang and Chan 2010) or the concrete use of sensors and other IoT technology (Pang et al. 2015).

3.2 Addressed challenges

75 articles deal with the subject of system integration and questions of interoperability. Articles tagged with the keywords system integration and interoperability cover technological issues like formats and standards (e.g. XML and RDF) as well as architectural approaches (Mezgár and Rauschecker 2014; Moisescu and Sacala 2014). In contrast, data integration articles mainly refer to logical challenges of the integration of distributed or unstructured data. Furthermore, 17 articles highlight the need for cross-company cooperation and 18 papers explicitly name challenges of information and knowledge sharing. Eventually, 12 articles address security and privacy issues (Fabian et al. 2013; Choi et al. 2013) and 8 articles discuss the demand for mobile solutions (Jones 2014).
3.3 Architectural building blocks

The architectures covered in the papers are usually distributed as well as federated and often involve some kind of platform or middleware for the integration and coordination of the actors (Nguyen and Ngo 2014; Gama et al. 2012). In 8 articles (mostly Service-oriented approaches), the role of such a middleware is comparable to a marketplace where the actors discover and negotiate about their cooperation (Menychtas et al. 2012; Bui et al. 2006).

71 articles discuss Cloud Computing. Most of them refer to Cloud Computing as an enabling technology that allows the provision of services on demand, according to the Everything-as-a-Service paradigm (Banerjee et al. 2011). The concept of Web Services is also seen as a major technological driver that often entails a vision of a (cloud-based) service ecosystem (Castejon et al. 2011; Smit et al. 2013). Another important technological term is infrastructure that encompasses aspects of underlying platform and network issues (Gama et al. 2012; Fogliata and Mussini 2008). The Semantic Web is often mentioned in connection to the integration of open and linked data (Konstantinou and Spanos 2015) as well as Software Agents which are mainly used for integration and coordination in dynamic environments (Maximilien and Singh 2005; Maciel et al. 2015). Finally, sensors and their role in IoT applications (McGrath and Scanaill 2013) as well as federation approaches for classical DWs (Golfarelli et al. 2010; Dymek et al. 2015) are addressed by 15 and 11 articles respectively.

The core insights from the literature review are therefore the identification of the link between IoT and federated BIA ecosystems, highlighting integration and knowledge-related challenges as well as the acknowledgement of solutions based on platform and service ideas.

4 Cases from the Realm of Advanced Manufacturing

In order to gain further insights into the subject in discussion, two case studies were conducted. As the literature review suggests, relevant applications can be expected in the area of the application of IoT and the “Industrial Internet”, in particular in the domain of (Advanced) Manufacturing, whereas the adjective “advanced” signifies the exploitation of IT and Internet technologies (Chung and Swink 2009). For this reason, two manufacturing companies with Advanced Manufacturing / Industrial Internet initiatives were selected.

Methodologically, we followed Yin (2013) who suggests case studies as an explorative research tool and provides a systematic approach to design, conduct and analyze a case. There has been a long term cooperation with both companies that provided a rich backdrop of information. The core results come from semi-structured on-site interviews with representatives responsible for the initiatives (Myers and Newman 2007). The interview guideline was structured according to the framework discussed above with added levels of detail regarding the architectural layers, the concrete data handling challenges, the IoT application itself, and the level of integration. The durations of the interviews were 120 and 90 minutes respectively. Both interviews were fully transcoded and iteratively coded using an open-coding approach (Miles and Huberman 1994).

4.1 Case 1: Cross-site ecosystem for Advanced Manufacturing

The company in the first case is a leading supplier for high-end transmission systems with about 200 globally distributed production sites and more than 50,000 employees. It was decided to bundle a variety of local Advanced Manufacturing and Industrial Internet projects under the umbrella of a company-wide strategy. A responsible Industrial Internet unit was formed that includes members from the executive level. Data integration and analysis are considered core parts of the pursued strategy.

The conceived and implemented applications particularly aim at achieving a traceability of objects across sites as well as root cause analysis when problems arise and that aim at an increased quality and reliability (esp. with the help of predictive maintenance). “In case a problem arises, our customers want to know if we can guarantee availability and if we can fix the root cause […] That is a
competitive advantage in our industry!”. Further objectives were more flexible material flows and in parts the provision of smart services. An overview of the situation of the company and its BIA ecosystem is given in Figure 3.

In order to harvest the benefits of the necessary analytical and managerial applications, data from novel “cyber-physical systems” (CPS), i.e. smart production and logistics machinery equipped with digital sensors and Intranet/Internet access, and heterogeneous data formats have to be integrated across sites and with data and documents from other business units, esp. research and development, and service and distribution. The forceful diffusion of CPS comes from the realization that “variability, dynamics, and individual customer requirements cannot be handled with flexible hardware and mechanics in particular”.

Due to the heterogeneous nature of the supported units, as well as the data volume and velocity of the newly collected CPS data, it was decided to follow a federated approach with both local and central data storage and analysis sub-systems. There is a variety of data analysis applications on machine, plant, and company levels. The amount of structured production data per plant is currently already at 6 TB for an average plant. This is increased dramatically by the inclusion of more and more semi-structured data: “For a sub-aspect on the subject ‘quality’, we came to an additional volume of 3 TB for one plant alone. [...] And I wished we had one quality system only in the group. But it will probably never come to that.” The monitoring and root-cause-analysis requirements increase the challenges: “We have products that deliver 3,000 parameters that we have to trace and analyze in a really short time frame as soon as a certain threshold is met.”

Currently, only data on events and developments (e.g. exceptions, averages) is shared across sites. The linking pins between the shop-floor and the higher order systems are formed by Manufacturing Execution Systems (MES) (MESA 1997; Kletti 2007) that combine functionality for production steering, data collection, (operational) reporting, and partly also more advanced analytics.

The core challenges of the approach result from bringing together heterogeneous divisions, business functions, and sites. This particularly comes with issues regarding data and system integration and the interoperability of the systems. Also, unavailable meta data is an issue (“E.g. I get a temperature value from a sensor: Of which machine, at which state of operation, in which parts?”). Also, the CPS bring new data quality challenges, e.g. due to failing sensors, or erroneous measurements.

Relevant solutions stem from the enforcement integrated standards, with a strong emphasis on the identification of objects by utilizing various automatic identification techniques. The central Industrial Internet unit primarily sees itself in the role of a governance unit that coordinates syntactic and
semantic standards and realizes overarching applications ("currently we are still a humungous pick-and-mix store with all sorts of coding and identification systems"). For the future, it is also conceived to centrally provide data analysis personnel and services from this unit, i.e. Data Scientists.

Currently, the Advanced Manufacturing production, logistics, and service data is kept separately from the “classical” administrative DW; the integration of these areas is on the strategic roadmap of the company however, as it promises to link technical situations with business indicators.

The two core insights from this case are:

1) Industrial Internet applications are indeed core drivers of federated BIA ecosystems, esp. when they are intended to a) span heterogeneous sites and business functions and b) come with high-volume machine data. In fact, the wider the approach and the more data is processed, the stronger the centrifugal forces that tear a centralized BIA approach apart.

2) The need for an overarching control of standards and formats to provide “classical” DW functions for meta data management and data administration in a federated solution.

4.2 Case 2: Cross-company platform for Machine Services

The company in the second case is a leading machine tool manufacturer that employs more than 10,000 people at more than 10 sites. This company also established an Industrial Internet unit. Unlike the first case, this unit is reporting to research and development. The reason for this stems from the fact that sensor technology, Internet integration, and analytic functionality are considered central features of the own products. The concepts for integrating the respective technologies are entering the product design in the very early phases of the product lifecycle. While IoT approaches are also used in-house, the focus here is more on the post-sales stage.

The new data gathering capabilities are particularly used for offering Predictive Maintenance services to customers. Examples are laser-cutting machines with special sensors for detecting lens abrasion. If a lens degrades in quality, the machines first counter the resulting effects with recalibration and later order in-time replacements, thus sparing the customer both waste and downtime. Respective machine-based services are heavy on data volume and velocity with the core data processing being executed on-site.

The idea of making the machines “smart” in order to provide new services has led to the conception and implementation of a customer-spanning platform that utilizes Cloud Computing technology for analytical services. Given the customer requests it, the analysis is controlled remotely by the machine tool provider under consideration of both customer and machine tool provider data. This particularly allows to conduct cross-company analysis e.g. on machine reliability and efficiency – across all participating companies willing to share their data in an anonymized fashion: "Of course I can execute reporting and analysis services [with the platform] for a single unit. But I can also say: In comparison with similar companies, your efficiency lies at 90% or 110%." The next planned steps are the integration of third-party machines, services for product flow steering that go beyond the individual machine as well as the provision of data integration and analysis services that include material suppliers and customers of customers. The chosen approach is depicted in Figure 4.

There is a strong transformational side to this IoT application: The roles of both the machine tool manufacturer and its customer change drastically, with the machine tool manufacturer moving into the direction of a supply-chain integrator. The long term vision are services for “machine capacity on demand” – based on BIA features.

Even more than in case 1, data volumes are considered to issues that prevent a singular central data store. Due to restrictions on the data capturing and storage side, a machine analysis is currently often limited to sensor data from the last couple minutes for immediate data processing.
Figure 4. BIA-Ecosystem of Case 2

There would be significant benefits if the respective data histories would be stored, esp. for reproducing the conditions leading to a failure or for a detailed comparison of issues and solutions across machines:

“On the lowest level, [...] the data is generated with 1 KHz [with on average 200 sensors per machine] [...] and everything related with geometry and technology-commands generates indefinite amounts of data! At first, this is relevant in real-time to track the state of the process and the machine. Actually, there is a need to track this data over time in order to analyze problems, e.g. in case of a crash or an accident [...]. Currently we are not doing that because of the data volume”. The company is currently working on applying In-Memory- and Big-Data-solutions to deal with these issues. It is already evident, however, that the relevant data volumes cannot be transferred to and stored in the Cloud, thus again leading to a federated architecture. Besides, security is a relevant concern here, as security breaches on machine steering level have consequences that go beyond the monetary aspect – consequences here can have an impact on the life and limb of the operating personnel.

The second case corroborates the two insights from the first case, particularly regarding the role of IoT as a driver of federated BIA solutions. Beyond that, it suggests four further points:

1) The application of a Cloud platform is a relevant architectural building block particularly for cross border BIA solutions.
2) There can be a necessity for keeping historical data for high-volume IoT data.
3) The role of BIA can change from a supportive one to a product feature that immediately contributes value and becomes an essential part of the core business processes. This further raises the bar for the design, monitoring, and administration of the BIA features.
4) A tight management, securing, and governing of the cross-application environments is essential.

In conclusion, a federated solution not only has to fulfill the basic Inmon DW characteristics (with the integration aspect being limited to shared data items) but also requirements that surpass those of a classical DW, esp. security and cross-border compliance.

5 Discussion: New challenges for BIA data management

A core result of the presented studies is the identification of a link between the application domain of the IoT and decentralized, federated approaches – a domain in which several forces drive an increasingly distributed data collection and analysis. The development leads from the DW over MES
(case 2) and machines (case 3) up to individual components or even sensors. Additionally, separated intra-company system landscapes are merged into cross-company digital ecosystems both on the technological and on the business side. This, however, comes with new requirements on the architectural side, e.g. when it comes to a secure data sharing. The well-known BIA challenges of data integration and interoperability approach become even more demanding. Both case studies strongly support these results.

**Challenge 1:** The evolution towards BIA ecosystems require methods to integrate data and systems across sites, business units and enterprise borders.

These developments also come with a tremendous increase in the number of involved systems. On the one hand, each participant adds new systems. On the other hand, the number of systems in general is increasing due to the inclusion of Big Data components, smart machines, and/or a multitude of sensors. The variety of scope, analytical capabilities (querying, data models etc.) and other characteristics further exacerbate the integration issues. Seemingly simple components like sensors come with only a limited set of functionality and meta data, challenges regarding availability or a restricted lifetime as well as new data quality issues (cf. case 1). The new system types also increase the dynamic character of the scenarios. In contrast to contemporary federated DW approaches, future architectures need to rapidly integrate and release systems (e.g. new actors in a site-crossing manufacturing process with an integrated machine tool provider).

**Challenge 2:** The number of participating components is increasing due to manifold new technologies, sub-systems, and components.

**Challenge 3:** Participating components can heavily differ in type, quality, and analytical capabilities.

**Challenge 4:** Participating components can quickly appear or vanish.

But not only do the types and composition of systems lead to new challenges, the data that is generated in such decentralized structures itself poses one. As both cases illustrate, the data generated in machines and sensors quickly exceeds the limits of a traditional DW. As the first case illustrates, new sources like sensors embody complex structured or unstructured content need to be processed, particularly when Industrial Internet solutions span the complete product life cycle and encompass systems for CAx, Product Lifecycle Management, and Quality. Often, this by default only works locally. However, DW requirements regarding data consistency and data quality remain. The same holds true for keeping a data history as Case 1 suggests. This leads to “local DW and Big Data” requirements.

A federated BIA ecosystem also needs the features known from DWs for data and meta data administration, data security, and data governance enforcement. This is salient in the first case but can be implicitly derived from the second case as well. In cross-company scenarios, security and governance become even more important.

**Challenge 5:** Large amounts of complex data has to be stored and evaluated quickly with a need for data storage, historization, and preparation at the or close to the data source.

**Challenge 6:** Despite the distributed nature of distributed IoT solutions, DW features for meta data and data quality management, data consolidation, and administration remain relevant.

On top of this, many emerging applications make use of the continuous stream of data to enable real-time decision making, e.g. for Predictive Maintenance or ad-hoc root cause analysis across processes. In order to enable such scenarios, there is a need to quickly load, integrate, and process data across distributed data sources.

**Challenge 7:** The demand for real-time decision making requires an immediate processing of heavily distributed data sources.

Current BIA architectures are not designed to deal with the combination of these seven challenges. One reason for this is that they are mainly conceived for solutions that can be handled in a central
While the “classical” DW data consolidation approach is valid for most non-real-time, aggregated reporting scenarios, it falls short of meeting the relevant requirements for distributed IoT solutions – even when realized with Big Data technologies. When projecting the discussed trends into the future, this becomes even more pressing – with an increasing number of individual sensors and embedded devices, increasingly spread across enterprise borders, the diffusion of Cloud scenarios, and a growing potential of cross-site and -company analytics.

### 6 Architecture vision

Figure 5 illustrates an architecture vision that aims at bridging the chasm between classical DW approaches and the challenges 1–7 as observed in the IoT / Advanced Manufacturing cases. It introduces the idea of moving parts of the DW functionality into self-contained analysis-ready pieces of information, e.g. located at an MES, a machine, or even an individual sensor. In fact, each entity would become a self-contained mini-DW, ready to be assembled ad-hoc to larger virtual DW components. This way, each subset with raw data is turned into what we call an Analytic Atom.

**Figure 5. Envisioned platform with Analytical Atoms**

The atom approach brings readiness for quick local analytics as demanded by challenge 5. Considering the challenges 1, 3, 4, and 6, each Analytic Atom needs to come with its own change history, include the subject area context, alternative versions as well as other meta data like spatial and process position. Neglecting such data would impede both the possibilities for an agile data integration and administration across units and enterprise borders. In addition to this, the atom needs to be able to conduct rapid, but reproducible and versioned data refinement and integration routines to meet the challenges 1, 2, and 6. By adding views, an Analytic Atom can furthermore behave in a polymorphic fashion and thereby better support cross-organizational BIA scenarios with differing data preparation for the local machine, the shop-floor / MES, the site, the company or the ecosystem level (e.g. when a supplier or a tool provider accesses a machine). Eventually, its actual data management needs to be kept transparent so that it can provide analytical capabilities in any extent and evoke Big Data components when needed (cf. challenge 5).

To deal with challenge 6, we propose a platform to manage data persistence, overarching meta data and governance rules. This platform can incorporate ideas discussed in section 2 and 3, especially by applying Cloud technologies for scalability and SOA technologies for dealing with interconnectivity (e.g. as proposed by Pospiech and Felden 2013). Ideally, the platform would be positioned between...
the data sources / source systems and services for data persistence. This way, the sources can act as independently as possible, while the platform automatically stores versions of the data, provides interfaces to other systems or adds meta-data e.g. based on the source system or an ETL tool applying changes. Provisioned via the platform and drawn from the Cloud, the Analytical Atom would gain access to functionality for data transformation and consolidation like format conversions, e.g. for currencies (as traditionally done in a DW), physical units (for IoT applications) or time-based data. Furthermore, a Cloud approach allows querying functionality and interfaces to higher-order administrative systems e.g. for Enterprise Application Management or IT Service Management suites.

The purpose of the platform is to bring the atoms together, coordinate their interplay, and administer the rules they follow. In a way, the platform provides ad-hoc virtual DWs that can be applied at the object, the site or the company level but allow analyses with combined data from various sites and business functions (as in case 1) or across companies (as in case 2).

With this approach, the architecture would be ready for agile scenarios that are still controlled and administrated according to defined governance rules. The challenge in this context is to provide as much functionality as possible in an automatic fashion so that the effort for maintaining the distributed solution is not higher than for a single DW. Another major challenge is to provide all this functionality while still satisfying performance and latency requirements. To enable real-time decision making with distributed data, as stated in challenge 7 (and complementing the local decision support at the atoms), the platform may also incorporate an additional speed lane with adequate streaming technology similar to the idea of Lambda Architectures (Marz 2015).

7 Conclusion and Outlook

Our findings from both the literature review and the two case studies strongly suggest that there are application areas for BIA where the current central DW approach is not suited and where it cannot be mitigated with established remedies known from the DW domain, e.g. quick-and-dirty EII solutions, data lakes, or In-Memory-BI. This is particular the case for IoT and Industrial Internet-based BIA applications in which relevant data pools cannot be merged easily into a central DW – for technical, legal or simply effort-related reasons. Moreover, a simple EII approach which merges raw data would also fail due to the loss of data history and a lack of meta data. The need for a well-governed and consistent approach is not diminished. In fact, for some aspects (e.g. data access and security) these requirements are even more pressing. With more and more distributed and partly mobile IoT data sources and integrated services on top of them, the BIA landscape becomes a mosaic of quickly moving pieces. To counteract this, we propose to move DW functionality to the individual data sources and build “mini DWs” that are “analysis-ready” and that form building blocks for ad-hoc virtual DWs. We call those Analytical Atoms. We also propose to use a partly Cloud-based federation platform for realizing this concept.

With respect to limitations, it needs to be acknowledged that the presented research is only a first step that covers the outlines of a new frontier for BIA. Both the literature and the cases reflect the early stage of the technological development. In both cases, the Advanced Manufacturing initiatives were relatively young. The projected situation of an utterly distributed data analysis landscape is not yet reality here. Given the visions in the Advanced Manufacturing sector that involve self-organizing factories which are geared towards a mass customization where each product is unique (lot size 1), the developments are indeed moving in the envisioned direction. Besides this still shaky empirical ground, the architecture vision itself is not yet realized and evaluated. This needs to be done in follow-up design-science oriented work with prototypes and an evaluation on real scenarios.

Nevertheless, the results provide a contribution to both the IoT and the BIA body of knowledge, particularly by the identification of IoT as a relevant domain for federated BIA architectures and by delineating the particular challenges that come with this development. Furthermore, the architectural vision of Analytical Atoms is a starting point for further research.
References


