Analytics for the Internet of Things: A Survey

EUGENE SIOW, University of Southampton
THANASSIS TIROPANIS, University of Southampton
WENDY HALL, University of Southampton

The Internet of Things (IoT) envisions a world-wide, interconnected network of smart physical entities. These physical entities generate a large amount of data in operation, and as the IoT gains momentum in terms of deployment, the combined scale of those data seems destined to continue to grow. Increasingly, applications for the IoT involve analytics. Data analytics is the process of deriving knowledge from data, generating value like actionable insights from them. This article reviews work in the IoT and big data analytics from the perspective of their utility in creating efficient, effective, and innovative applications and services for a wide spectrum of domains. We review the broad vision for the IoT as it is shaped in various communities, examine the application of data analytics across IoT domains, provide a categorisation of analytic approaches, and propose a layered taxonomy from IoT data to analytics. This taxonomy provides us with insights on the appropriateness of analytical techniques, which in turn shapes a survey of enabling technology and infrastructure for IoT analytics. Finally, we look at some tradeoffs for analytics in the IoT that can shape future research.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Information systems** → *Data analytics*; • **Networks** → *Network architectures*; *Cyber-physical networks*;

Additional Key Words and Phrases: Internet of things, data analytics, cyber-physical networks, big data

ACM Reference format:
Eugene Siow, Thanassis Tiropanis, and Wendy Hall. 2018. Analytics for the Internet of Things: A Survey. *ACM Comput. Surv.* 51, 4, Article 74 (July 2018), 36 pages.
https://doi.org/10.1145/3204947

1 INTRODUCTION

The Internet of Things (IoT) has been gaining momentum in both the industry and research communities due to an explosion in the number of smart mobile devices and sensors and the potential applications of the data produced from a wide spectrum of domains. In their 2013 report, McKinsey note a 300% growth in connected IoT devices in the past five years and rate the potential economic impact of the IoT at $2.7 trillion to $6.2 trillion annually by 2025 [126]. These figures grew to $4 trillion and $11 trillion in 2015 [125]. A study of Gartner's 2010 to 2017 hype cycle reports, which we aggregate in Figure 1, shows the advent of the IoT, steady growth, expansion, and creation of new technology areas like the IoT platform. Another interesting technology that exceeds the IoT in momentum on the hype cycle is that of big data, which the IoT serves as a source and sink of.

Authors' addresses: E. Siow, T. Tiropanis, and W. Hall; emails: sboonlin@dso.org.sg, t.tiropanis@southampton.ac.uk, wh@ecs.soton.ac.uk.

https://doi.org/10.1145/3204947

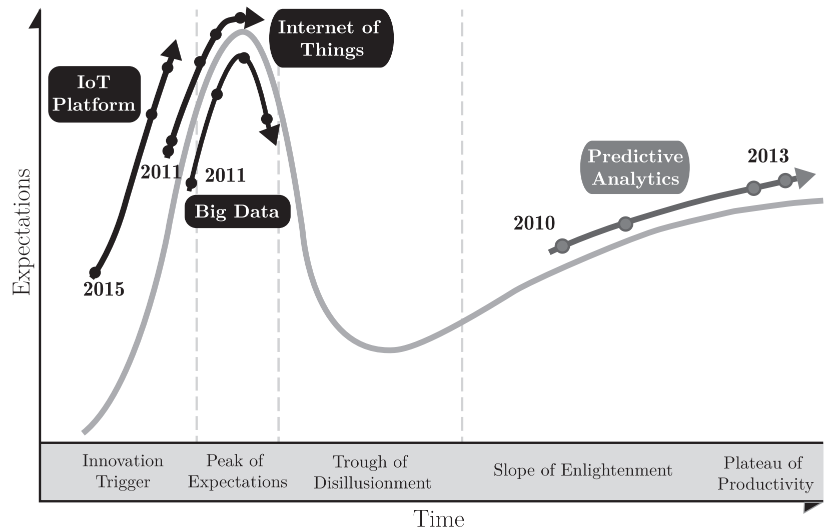

Fig. 1. Aggregated Gartner hype cycle of technologies from 2010 [61, 145, 152–154, 169–171].

Big data is data that are *too big* (volume), *too fast* (velocity), and *too diverse* (variety) [122]. In the context of the IoT, we see an example of volume in the DEBS 2014 Grand Challenge [218], where data from 40 houses with smart plugs produced 4 billion events in a month [60], given that a 2011 census showed that there were 26.4 million households in the United Kingdom [142], the projected data size of 2.64 quadrillion (short scale) per month if every house had a meter, is a good example of *too big* data. In the IoT use cases of intelligent transportation systems [143, 195] and telecommunication, data streams can come in *too fast* for processing, representing a data *velocity* problem. Finally, *too diverse* is the catchall term used to describe the presence of heterogenous data sources in the IoT that make it difficult for existing tools to analyse them. In a 2014 survey of data scientists, 71% interviewed said that analytics is becoming increasingly difficult due to the variety and types of data sources [147]. An example is in the personal health care use case of the IoT [141], where unstructured textual electronic health records, connected mobile devices, and sensors [11] all add to the variety problem.

Analytics is the science or method of using analysis to examine something complex [144]. When applied to data, analytics is the process of deriving (the analysis step) knowledge and insights from data (something complex). The evolution to the concept of analytics we see today can be traced back to 1962. Tukey first defined data *analysis* as procedures for analysing data, techniques for interpreting the results, data gathering that makes analysis easier, more precise, and accurate, and finally, all the related machinery and statistical methods used [192]. In 1996, Fayyad et al. published an article explaining Knowledge Discovery in Databases (KDD) as "the overall process of discovering useful knowledge from data," where data mining serves as a step in this process—"the application of specific algorithms for extracting patterns from data" [59]. In 2006, Davenport introduced analytics as quantitative, statistical, or predictive models to analyse business problems like financial performance or supply chains and stressed its emergence as a fact-based decision-making tool in businesses [50]. In 2009, Varian highlighted the ability to take data and "understand it, process it, extract value from it, visualise it and communicate it," as a hugely important skill in the coming decade [197]. In 2013, Davenport introduced the concepts of Analytics 1.0, traditional analytics, 2.0, the development of big data technology, and 3.0, where this big data technology is integrated agilely with analytics, yielding rapid insights and business impact [51].

Table 1. Chronological Summary of Previous Surveys in the IoT, Big Data, and Analytics

Year	Reference	*I*	*B*	*A*	Description
2010	Atzori et al. [16]	✓			Vision, Apps
	Sharma et al. [182]			✓	Business Analytics
2012	Miorandi et al. [133]	✓			Vision, Challenges
	Barnaghi et al. [21]	✓			Semantics
	Chen et al. [36]			✓	Business Analytics
2013	Sagiroglu et al. [173]		✓		Problems, Techniques
	Vermesan et al. [199]	✓			Vision, Apps, Governance
2014	Perera et al. [151]		✓		Context-aware
	Zanella et al. [216]	✓			Smart Cities
	Xu et al. [208]	✓			Industries
	Zhou et al. [217]		✓	✓	Big Data Analytics Challenges
	Kambatla et al. [103]		✓	✓	Big Data Analytics Trends
	Chen et al. [38]	✓	✓	✓	Big Data Analytics
	Stankovic [188]	✓			Directions
2015	Al-Fuqaha et al. [7]	✓	✓		Protocols, Challenges, Apps
	Granjal et al. [72]	✓			Security Protocols, Challenges
2016	Ray [166]	✓		✓	IoT Architectures
	Razzaque et al. [168]	✓			Middleware for IoT
2017	Akoka et al. [6]		✓		Big Data Research Trends
	Lin et al. [116]	✓			Fog Computing Architecture
	Farahzadia et al. [58]	✓	✓		Middleware for Cloud IoT
	Sethi et al. [181]	✓			IoT Architectures, Apps

Legend: *I*=IoT, *B*=Big Data, *A*=Analytics.

To better understand each of these areas, the IoT, Big Data and Analytics, and their intersection, we look chronologically at the existing reviews and surveys on these topics. This will help to establish the need for our review from the new dimension of analytics on the IoT especially in big data scenarios. A summary of the reviews is shown in Table 1.

In 2010, Atzori et al. [16] survey the vision of the IoT, the enabling technologies, and potential applications while identifying three perspectives: Things, Semantics, and Network. Sharma et al. [182] study analytics applications in the industry and propose a framework of how business analytics can be applied to processes for organisations to gain a sustainable, competitive advantage.

In 2012, Miorandi et al. [133] survey the IoT mainly from the perspective of the key issues and research challenges and some initiatives going on to address them. Barnaghi et al. [21] look at developments in the semantic web community, analysing the advantages of semantics but also highlighting the challenges they face and review work on applying semantics to the IoT. Chen et al. [36] study, using bibliometrics, some of the key research areas in business intelligence and analytics and some application areas, and they propose a framework to classify them.

In 2013, Sagiroglu et al. [173] gave an overview of the big data problem, methods to handle the big data, analysis techniques, and challenges. Vermesan et al. [199] look at the vision, applications, governance, and challenges of the IoT and some proposed solutions like semantics.

In 2014, Perera et al. [151] presented a study of context-aware computing and discussed how it can be applied to the IoT. Zanella et al. [216] survey the enabling infrastructure and architecture for the Internet of Things in an urban, connected, smart city scenario, while Xu et al. [208] review the development of IoT technologies for industries. Zhou et al. [217] discuss the challenges brought to data analytics by big data from the perspective of various applications while Kambatla et al. [103] discuss trends with a focus on hardware and software platforms, virtualisation, and application scopes for analytics. Another big data survey is done by Chen et al. [38] who look at challenges and work done from each stage of "data generation, data acquisition, data storage, and data analysis." They also look at applications of big data briefly, where one such area is the IoT. Finally, Stankovic [188] proposes a set of research directions and considerations for future research on the IoT.

In 2015, Granjal et al. [72] surveyed existing protocols for protecting communications on the IoT, comparing against a set of fundamental security requirements, and they highlighted the open challenges and strategies for future research work. Al-Fuqaha et al. [7] focus on giving a thorough summary of protocols for the IoT and how they work together for applications in big data scenarios.

In 2016, Ray [166] surveyed domain-specific architectures for the IoT providing a brief summary of whether cloud platforms in the IoT support data analytics. Razzaque et al. [168] survey middleware platforms for the IoT against a set of comprehensive service and architectural requirements.

In 2017, Akoka et al. [6] performed a systematic mapping study, a method for structuring a research field, to classify big data academic research and identify trends in the research. Both analytics and the IoT were identified as popular topics. Reviews by Lin et al. [116] and Farahzadia et al. [58] focus on specific IoT research areas of fog computing architectures and middleware for cloud computing platforms. Sethi et al. [181] take the approach of surveying IoT architectures, protocols, and applications, which help them organise a taxonomy of IoT research.

One can see that the vision of the IoT through these surveys is still very much about interconnecting physical objects with protocols, however, the introduction of Big Data and Analytics has meant that there has been a broadening of focus from communications technologies to applications with impact, scalability, and utilising context within cross-domain use cases like the smart city while also coalescing around fog computing and edge technologies, middleware platforms, and the cloud. Information rather than data is increasingly envisioned as the new language of the IoT, while infrastructure and enabling technologies have shifted toward dealing with Big Data use-cases with high scalability or within distributed systems.

Given the traction of big data analytics in the industry and the IoT's potential to become a "dominant source" of big data [39], while also a consumer of insights and optimisation drawn from analytics, we foresee that researchers will be looking to understand the process of deriving analytical insights from the IoT. This is further justified by the argument of Akoka et al. [6] that "data of IoT is useful only when analyzed." As we have noted in our chronological study of previous reviews, this particular combination of areas, with a focus on IoT analytics, to the best of our knowledge, has not been explored in depth. The contribution of this article is then to:

(1) review IoT analytics applications and research from a variety of domains,
(2) propose a classification and taxonomy for IoT analytics to guide future work, and
(3) review the enabling infrastructure for analytics in the context of big data and examine the tradeoffs to shape research directions.

The methodology used and organisation of the rest of the article is explained in Section 2.

2 METHODOLOGY AND ORGANISATION OF ARTICLE

Section 3 starts by introducing the IoT vision and application domains, highlighting how this motivates this article, which is then followed by the main survey content of the article. The approach employed for the survey follows that of an evidence-based systematic review [108]. First, two research questions (RQ) were framed:

RQ1 What IoT analytics research/applications are being published?
RQ2 What enabling infrastructure is required for big data IoT analytics applications?

Next, we employed an approach of identifying relevant articles through search on the Web of Science platform that indexed an extensive list of multi-disciplinary journals and conferences across multiple databases. The search criteria included the keywords "`big data`" or "`analytics`," filtered by "`internet of things`." From 2011 to 2015, 460 articles were retrieved. This was updated with 311 articles from 2016 and 2017 when the paper was revised.

The articles were further screened manually following an inclusion criteria that mandated they (1) were from original research, (2) described actual designs, implementations, and results, (3) applied analytics, and (4) served IoT use-cases. The highest ranked six papers were chosen from each of five IoT application domains determined from IoT literature, forming a high-quality pool of 30 papers according to the systematic review method. This addressed RQ1 and is presented in Section 4. The ranking was decided proportionately by the number of citations and a qualitative score from 0 to 5 of the technological complexity and completeness of the application (mitigating recency bias).

This understanding of IoT applications was combined with business analytics literature, which has successfully drawn insights from data to optimise business processes, to propose a classification for analytics in Section 5. This classification will help us to better define and target research through an IoT analytics taxonomy as part of the summarisation step of a systematic review.

Finally, we go on to review the current state-of-the-art in IoT infrastructure in Section 6 that answers RQ2. We used the survey and applications publications previously retrieved on the IoT and identified, by manual inspection, groups of work in cloud, middleware, distributed, and fog computing and expanded the search through these keywords to retrieve relevant articles for IoT scenarios as part of the "interpreting the findings" step. Our goal was to consider analytics infrastructure from the perspective of data generation, collection, integration, storage, and compute.

The rest of the article consists of research challenges (Section 7) and a conclusion (Section 8).

3 THE INTERNET OF THINGS VISION AND APPLICATION DOMAINS

3.1 IoT Definition and Common Vision

Both the European Commission and the UK Government Office of Science have a similar vision of the IoT as "a world in which everyday objects are connected to a network so that data can be shared," greatly impacting society [57, 202]. The International Telecommunication Union (ITU) calls the IoT "a global infrastructure for the information society, enabling advanced services by interconnecting things based on existing and evolving interoperable information and communication technologies" [91] and from a broader perspective, "a vision with technological and societal implications," which draws its language from a report by the World Economic Forum [206].

Common to each of these visions are four principles that are well-defined in IoT literature:

(1) The IoT exists at a global scale [113, 199, 200],
(2) consists of uniquely identifiable Things with sensing or actuating capabilities linked to the physical world [16, 110, 213],
(3) which are interconnected by existing or future technologies so that data can be shared [7, 133],
(4) and have potential for societal impact through advanced services [181, 188, 208].

The motivation of this article builds on the third and fourth principles to identify and understand how analytics can enable advanced services from shared and integrated IoT data. The goal then, from these findings, would be to develop various means to help determine what analytics need

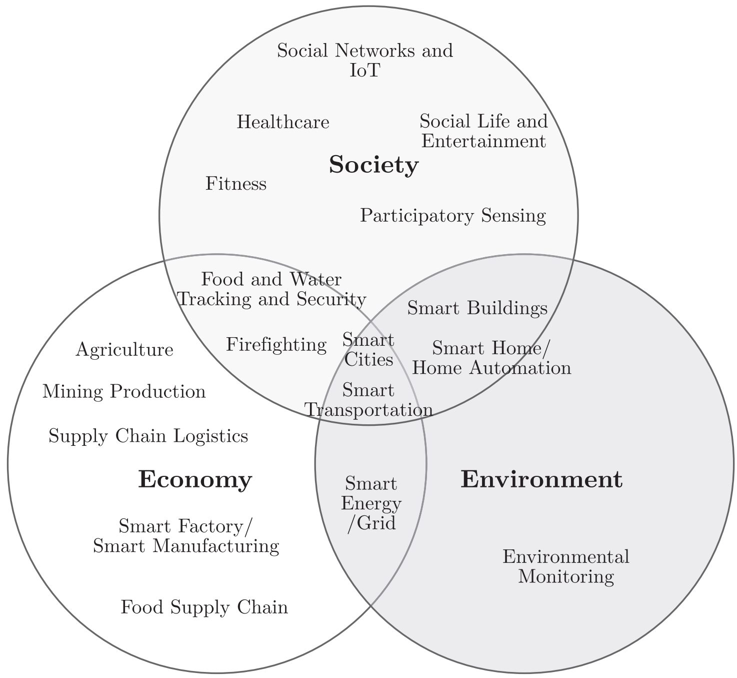

Fig. 2. Application areas from surveys categorised by impact to society, environment, and economy.

to be applied and what enabling infrastructure is necessary. First though, we need to define "advanced services." The next section builds on previous literature to define a set of advanced services domains that help organise the survey of analytics and ensure the article fulfils a broad coverage.

3.2 IoT Advanced Services and Application Domains

As the IoT develops, many more potential applications and use cases for the IoT will emerge, providing advanced services that offer positive externalities [85]. A range of advanced service application areas were elicited from each of the surveys describing applications from the 22 in Table 1 in Section 1. They were then classified under their impact to the themes of environment, society, and economy, which are the drivers of sustainable development used for analysing medium to long term development issues at a large scale [65]. Figure 3 shows the categorisation of the various application areas according to their economic, environmental and societal impact.

From these applications areas, a range of application domains, including health, transport, living, environment, and industry are used to group them, forming the hierarchical classification shown in Figure 3. Certain IoT research topics like Smart Cities [30], Smart Transportation [186], and Smart Buildings and Smart Homes [32], which impact multiple themes, are also listed.

This classification of advanced service application domains elicited from previous literature serves to advise, organise, and ensure the broad coverage of the following survey on IoT analytics applications and infrastructure in Sections 4, 5, and 6.

4 IOT APPLICATIONS WITH ANALYTICS

An important question to ask following our definition of the IoT and its vision is the advantage that connected "things" offer over isolated devices. For example, what is the benefit of deploying a

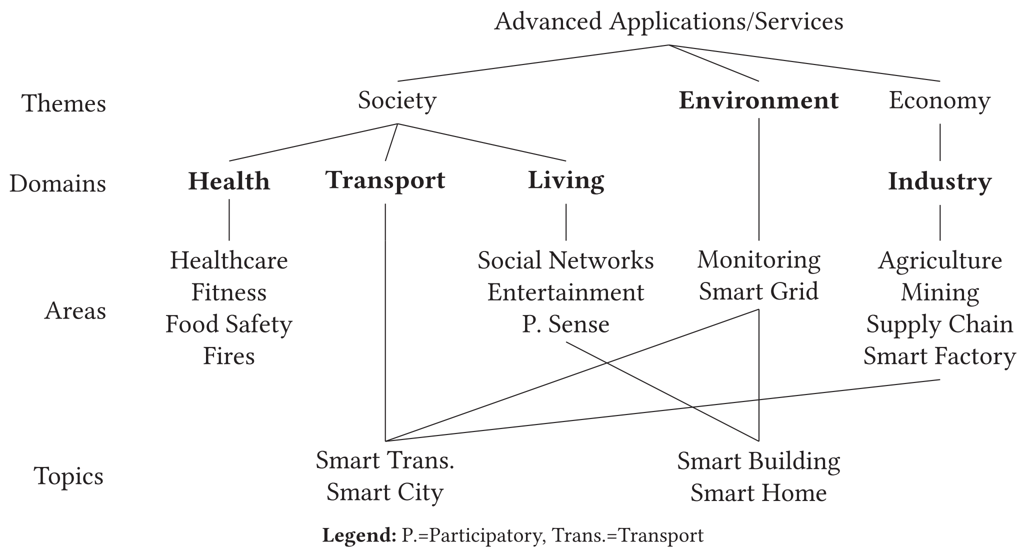

Fig. 3. Application themes, **domains**, and areas hierachy.

smart parking system as compared to having isolated sensors in a car park using visual signals of green or red on the ceiling to indicate whether a parking lot is empty or occupied? Analytics adds value to integrated data and context from the IoT, producing higher value insights. The analytics-powered smart parking system has a much wider observation space and also guides the user to the available parking lot efficiently, without human intervention, reducing traffic and pollution [17, 175].

Research publications of IoT applications that make use of analytics from 2011 to 2017 were surveyed and the top six based on the systematic review methodology (Section 2) from each application domain introduced in Section 3.2 is presented. This is described as follows and summarised in Table 2, which includes the analytics techniques employed, data sources used, and the currency of the data. Currency refers to whether analytics was applied mainly on historical or real-time data.

4.1 Health: Ambient Assisted Living, Neo-natal Care, Prognosis, Monitoring

Mukherjee et al. [136] review the use of data analytics in healthcare information systems. Two analytics applications are Ambient Assisted Living (AAL) [56] and neo-natal care. In AAL, rules are applied to IoT data collected from smart objects in the homes of elderly or chronic disease patients while advanced solutions take into consideration contextual information and apply inferencing using ontologies to give health advisories to users, update care-givers, or contact the hospital in emergencies. By analysing contextual knowledge in connection with physiological data and being sensitive and adaptive to parameters that vary less frequently, such systems are able to provide descriptive analytics to care-givers and a form of discovery analytics to detect anomalies to trigger emergency warnings. Neo-natal care involves the care of newborn babies where data mining is applied to multiple data streams to find relationships and patterns and to diagnose any possible medical conditions in infants who are not able to give the doctors verbal feedback.

Analysing the content of video data to aid the elderly and visually handicapped for AAL and navigation, respectively, is another IoT healthcare application of analytics [117].

In their work, Chen et al. [37] design a smart clothing monitoring system with visualisations of wearable sensor data through a mobile application for use cases like baby, elderly, and fitness

Table 2. Summary of Analytics Applications by Domains

Application	Data Sources	Technique	Currency
Health			
Neo-natal Care & AAL [136]	Sn+M_r	Data mining, Rules	R
AAL & Navigation [117]	Video	Video Analytics	R
Smart Clothing Monitoring [37]	Sn	Visualisation + ML	H
ECG Health Monitoring [86]	Sn	Watermark + Classifier	R
Prognosis [87]	Sn+M_r	Data mining	H
Wellness Recommendations [20]	Sn	Classifier + Rules	H
Transport			
Traffic Control [124]	Video	Video Analytics	R
Pedestrian & Car Detection [47]	Sn+Video	Video Analytics + CV	R
Behaviour & Traffic Prediction [99]	Sn	Visualisation + Model	H
Travel Routing [115]	Sn	CRF, A*Search	R
Smart Parking [82]	Sn	Model	R
Parking Anomaly Detection [155]	Sn	Self-Organising Maps	H
Living			
Cultural Behaviour [41]	Sn+S_m	Visualisation + Model	H
Police Situational Awareness [167]	Sn	Visualisation	H
Public Safety Monitoring [66]	Video	Video Analytics	R
Smart Building Heating [157]	Sn	Anomaly Detection	R
Memory Augmentation [75]	IoT Data	Data mining	H
Wearable Lifestyle Monitor [138]	Sn	Anomaly Detection	R
Environment			
Disaster Detection & Warning [178]	Sn+S_m	Anomaly Detection	R
Urban Disaster Storytelling [210]	S_m	Data mining	R
Wind Forecasting [135]	Sn	ANN	H
Energy Usage Recommendations [9]	Sn	ML + Rules	H
Energy Policy Planning [4]	Sn	Classifier + Models	H
Smart Energy System [64]	Sn	Data mining	H
Industry			
On Shelf Availability [196]	Video+Sn	Video Analytics	R
SCM Environment Control [140]	Sn+Traffic	CEP	R
SCM 4PL [172]	Sn	Ontologies	R
Floricultural SCM [198]	Sn+Traffic	CEP	R
Smart Farming [104]	Sn	CEP + Ontologies	R
Chemical Process Monitoring [42]	Sn	ML (ANN, Gaussian)	R

Legend: AAL=Ambient Assisted Living, ANN=Artificial Neural Network, CEP=Complex Event Processing, CRF=Conditional Random Fields, CV=Computer Vision, H=Historical, MDP=Markov Decision Process, ML=Machine Learning, M_r=Medical Records, R=Real time, SCM=Supply Chain Management, Sn=Sensors, S_m=Social Media.

monitoring. This data is also stored on a "health cloud" integrated with a machine-learning library for diagnostic and predictive analytics of medical conditions and users health trends, respectively.

Hossain and Muhammad [86] show how electro-cardiogram (ECG) and other healthcare data collected from wearable IoT devices and sensors can be watermarked to ensure integrity and sent to the cloud for analysis through feature extraction and classification with a support vector machine (SVM) in real time. Abnormal patterns are discovered and healthcare professionals alerted.

Analytics can also be applied in the form of prognosis, the science of predicting the future medical condition of a patient, to help healthcare professionals make more informed decisions [87]. Health indicators collected from sensors of a patient can be compared with data of similar patients and combined with domain knowledge and medical research to make conjectures.

Banos et al. [20] developed a digital health and wellness framework that collects data streams of IoT health data forming a "life log" for each user and includes descriptive analytics visualisations of activities. A human activity recogniser combines signal processing, SVM, and Gaussian Mixture Models to distinguish activities and recommends activities using rule-based reasoning.

4.2 Transport: Traffic Control and Routing, Pedestrian Detection, Smart Parking

Applying analytics on video content has a variety of applications in different fields. In their review paper, Liu et al. [117] looked at the latest technologies and applications of video analytics and intelligent video systems. Video analytics has been successfully applied in traffic control systems to detect traffic volume for planning, highlighting incidents and enhancing safety by enforcing traffic rules [124]. Another set of applications is for intelligent vehicles to assist the driver. Danner et al. [47] introduce their Precedent-Aware Classification (PAC) technique, which combines information from previously traveled routes and minimal classification features from sensors to computer vision analytics for pedestrian and car detection on constrained IoT platforms.

Jara et al. [99] derive insights about human dynamics by analysing the correlation between traffic, temperature, and time using IoT sensor data from the SmartSantander smart city testbed [176]. They apply visual analytics to understand and discover insights on human behaviour and use a poisson model to interpolate and predict traffic density. Liebig et al. [115] go further by prescribing good routes in travel planning using analytical techniques (a spatiotemporal random field based on conditional random fields [149] for traffic flow prediction and a Gaussian process model to fill in missing values in traffic data) to predict the future traffic flow and to estimate traffic flow in areas with limited sensor coverage. These were then used to provide the cost function for the A* search algorithm [80] that uses the combination of a search heuristic and cost function to prescribe optimal routes (provided the heuristic is admissible and predicted costs are accurate).

He et al. [82] develop a smart parking service that combines geographic location information, parking availability, traffic, and reservation information. The parking process is modelled as a birth-death stochastic process, which allows prediction and optimisation of parking availability. Piovesan et al. [155] describe the application of their unsupervised form of self-organising maps (SOM) clustering to the classification of parking spaces according to spatio-temporal patterns. This type of analytics automatically discovers outliers for sensor maintenance and usage anomalies.

4.3 Living: Cultural Behaviour, Public Safety, Smart Buildings, Memory Augmentation, Lifestyle Monitoring

Chianese et al. [41] describe a system for cultural behaviour analysis. They combine models and proximity evaluation algorithms to classify movement in museums from sensors with semantic enrichment from knowledge bases of cultural exhibits and social media of cultural tourism to analyse cultural behaviour using visualisations within an associative model.

Visualisation that taps the human cognitive ability to recognise patterns has also been employed by Razip et al. [167] in helping law enforcement officers increase their situational awareness. Officers are equipped with mobile devices that tap into crime data and spatio-temporal sensor data to show interactive alerts of hotspots, risk profiles, and on demand chemical plume models.

Additionally, there are public safety and military applications that apply video analytics in detecting movement, intruders or targets. The public safety use case is elaborated on by Gimenez et al. [66], where they discuss how given the big data problem of having huge amounts of video footage, smart video analytics systems can proactively monitor, automatically recognise and bring to notice situations, flag out suspicious people, trigger alarms, and lock down facilities through the recognition of patterns and directional motion, recognising faces, and spotting potential problems by tracking, with multiple cameras, how people move in crowded scenes.

Ploennigs et al. [157] show how analytics can be applied to energy monitoring used in heating for smart buildings. The system is able to diagnose anomalies in the building temperature, for example, break downs of the cooling system, high occupancy of rooms, or open windows causing air exchange with the external surroundings. Using a semantics based approach, the Building Automation and Control Systems (BACS) [15] could, from the sensor definitions, automatically derive diagnosis rules and behaviour of a specific building, making it sensitive to new anomalies.

Guo et al. [75] look at discovering various insights from mining the digital traces left by IoT data from cameras, wearables, mobile phones, and smart appliances. Resulting applications are life-logging systems to augment human memory with recorded data, real-world search for objects and interactions with people and a system to improve urban mobility systems by studying large-scale human mobility patterns.

Mukherjee et al. [138] present a fast algorithm for detecting anomalies and also for classifying high-dimensional data. These were tested with accelerometer data from a wearable personal digital assistant to recognise human activity in real time but can be generalised to other types of high-dimensional data. The importance of such algorithms in detecting anomalies and discovering patterns to classify activity from sensor data are analytical tools that form a basis for smart and intelligent devices and, in this example, for activity tracking and monitoring.

4.4 Environment: Disaster Detection & Response, Wind Forecasting, Smart Energy

Schnizler et al. [178] describe a disaster-detection system that works on heterogenous streams of sensor data. Their method includes Intelligent Sensor Agents (ISAs) that produce anomalies, low-level events with location and time information, e.g., an abnormal change in mobile phone connections at a ISA in a telecom cell or base station, a sudden decrease in traffic, increase in twitter messages, change in water level, or change in the volume of moving objects at a certain location. These anomaly events then enter Round Table (RT) components that fuse heterogenous sources together by mapping them to a common incident ontology through feedback loops that might involve crowdsourcing, human-in-the-loop, or adjusting parameters of other ISAs to find matches. The now homogenous incident stream can then be processed by a Complex Event Processing (CEP) [120] engine to complete the situation reconstruction by doing aggregation and clustering with higher-level semantic data, simulation, and prediction of outcomes and damage. The resultant incident stream can provide early warning, effecting early disaster response.

Xu et al. [210] also present a disaster-detection system targeted instead at urban disasters. They utilise social media events from multi-modal microblog posts (videos, images and text) to mine semantic, spatiotemporal, and visual information producing a story. This real-time story of urban emergencies unfolding serves to increase the situational awareness of emergency response teams.

Another environmental application is wind forecasting [135]. Data is collected from wind-speed sensors in wind turbines and an Artificial Neural Network is used on this data and historical data to perform the forecasting. This is useful for energy provision and planning.

Ghosh et al. [64] have implemented a localised smart energy system that uses smart plugs and data analysis to actively monitor energy policy and by performing pattern recognition analysis on accumulated data, spot additional opportunities to save energy. This resulted in saving on electricity bills especially by reducing the amount of power wasted in non-office hours from appliances, desktops, and printers.

Similar work by Alonso et al. [9] works on using machine learning and an expert system (rule-based) to provide personalised recommendations, based on energy usage data collected in Smart Homes, that help a user to more efficiently utilise energy. They go one step further to provide recommendations through predicting cheaper options by detecting similar patterns in big data collected from other homes. Ahmed [4] applies similar analysis on combined consumption data for use in organisations to help in energy policy planning. He develops a model to classify the energy efficiency of buildings and the seasonal shifts in this classification, and using more detailed appliance-specific data, forecasts future energy usage.

4.5 Industry: Supply Chain Management, Smart Farming, Chemical Process

Vargheese et al. [196] propose a system that improves shoppers' experience by enhancing the On the Shelf Availability (OSA) of products. Furthermore, the system also looks to forecast demand and provide insights on buyers' behaviour. A multi-tiered approach is employed, where sensors like video cameras, process video streams locally and analyse the products on the shelf, this data is then verified by other sensors like light, infra-red, and RFID sensors, and the metadata produced is sent to the the cloud to be further processed. In the cloud, this real-time data is combined with models from learning systems, data from enterprise Point of Sale (POS) systems and inventory systems to recommend action plans to maintain the OSA of products. The staff of the store are informed and action is taken to restock products. Weather data, local events, and promotion details are then analysed with the current OSA to provide demand forecasting and to model buyers behaviour, which is fed back into the system.

Nechifor et al. [140] describe the use of real-time data in analytics in a cold chain monitoring [2] process. Trucks are used for transporting perishable goods and drugs that require particular thermal and humidity conditions, sensors measure the position and conditions in the truck and of each package, while actuators—air conditioning and ventilation—can be controlled automatically. On a larger scale, predictions can be made on delays in routes and when necessary to satisfy the product condition needs, longer but faster routes (less congestion) might be selected.

Similarly, Verdouw et al. [198] and Robak et al. [172] examine supply chains—the integrated, physical flow from raw material to end products with a shared objective—and formulate a framework based on their virtualisation in the IoT. At its highest level, a virtual supply chain supports intelligent analysis and reporting. This is applied to floricultural and a Fourth Party Logistics (4PL) integrator, respectively, where business intelligence, data mining, and predictive analytics can provide early warning in case of disruptions or unexpected deviations and advanced forecasting about consequences of the detected changes when the product reaches destination.

In the above examples on product and supply chain management, we see a common theme of predictive analytics being employed to business processes. This predictive analytics is often powered by learning from data to discover models or through data mining for patterns in data. The effectiveness of these algorithms benefits from the big data of the IoT in providing a large observation space for discovering patterns and trends. Real-time data from sensors then provide the information required to immediately control actuators to rectify problems like products being out of stock on the shelf or conditions in trucks being unsuitable for perishable food.

Kamilaris et al. [104] describe the use of a Complex Event Processing (CEP) engine to discover significant events on semantically enriched data streams from sensors within two smart farming scenarios. One scenario included detecting the fertility of cows from temperature readings and other information on a dairy farm to suggest the best insemination timings. The other was to adaptively control the soil conditions for crop cultivation.

The chemical process industry deploys inferential industrial IoT sensors to process monitoring chains [42]. Some techniques applied by sensors include linear regression, artificial neural

Table 3. Summary of Application References by their Domains and Analytical Capabilities

Domain \ Capability	Descriptive	Diagnostic	Discovery	Predictive	Prescriptive
Health	[37, 136]	[37, 117]	[20, 86, 136]	[87]	[37]
Transport	[124]		[47, 99, 155]	[82]	[115]
Living	[41, 167]	[157]	[41, 66, 75, 138]		
Environment	[210]		[64, 178]	[4, 9, 135]	
Industry			[172, 198]	[42, 140, 198]	[42, 104, 196]

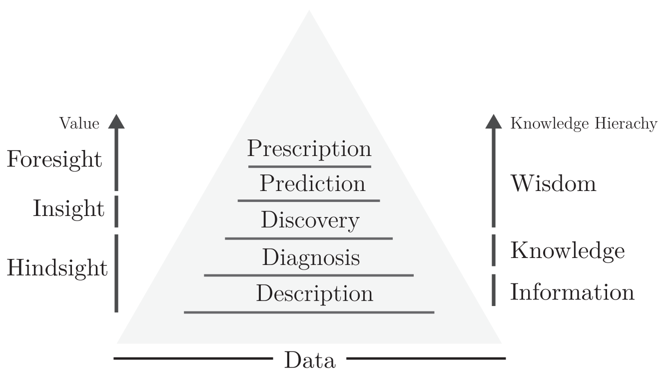

Fig. 4. Analytics and the knowledge and value hierachies.

networks (ANN), and Gaussian process regression, which predict variables using available process data. These predictions enable quality monitoring and advance control systems in plants to automatically react and prescribe process modifications "to prevent off-grade products."

5 TYPES OF ANALYTICS AND THEIR IMPORTANCE

Following the study of the current work in analytics in the IoT, we explore a classification of analytics that is applicable to these domains. We derive a categorisation of analytical capabilities from business analytics literature, which the term analytics comes from. Bertolucci et al. [26] propose descriptive, predictive, and prescriptive categories while Gartner [105], [33] proposes the extra category of diagnostic analytics. Finally, Corcoran et al. [45] introduce the additional category of discovery analytics. We build upon these to form a comprehensive classification of analytic capabilities consisting of five categories: descriptive, diagnostic, discovery, predictive, and prescriptive analytics. Each category is described in detail in Section 5.1, and we also summarise how each IoT application surveyed in the previous section is categorised in Table 3. Each application domain has applications that support multiple analytical capabilities. We also note that all the categories of capabilities are well-represented in the literature survey, while mature domains like the industrial IoT focus on high-value analytics.

Figure 4 looks at how each analytical capability fits within the Knowledge Hierarchy [24], which is a common framework used in the Knowledge Management domain. This categorisation of analytic capabilities enables us to establish what the aim of analysis is and allows us to relate to the vision of IoT deployment as often expressed in research roadmaps. The value of each capability is also highlighted in the figure. The knowledge hierarchy starts with data at the base, examples of which are facts, figures, and observations (e.g., the raw data produced by IoT "things"). Information is interpreted data with context, for example, temperature as represented by descriptive analytics: an average over a month or a categorical description of the day being sunny and warm.

Knowledge is information within a context with added understanding and meaning, perhaps possible reasons for the high average temperature this month. Finally, wisdom is knowledge with insight, for example, discovering a particular trend in temperature and projecting it across future months while providing cost-saving energy management solutions for heating a smart home based on these predictions. Each component of the knowledge hierarchy builds on the previous tier, and we can see something similar with analytical capabilities. To add a practical view from business management literature to our discussion, a review of organisations adopting analytics [112] categorised them as Aspirational, Experienced, and Transformed. Aspirational organisations were seen to use analytics in hindsight as a justification for actions, utilising the data, information, and knowledge tiers in the process. Experienced organisations utilised insights to guide decisions, and transformed organisations were characterised by their ability to use analytics to prescribe their actions, effectively applying foresight in their decision making process.

5.1 Five Categories of Analytics Capabilities

5.1.1 Descriptive Analytics. It helps us to answer the question, "What happened?" It can take the form of describing, summarising, or presenting raw IoT data that has been gathered. Data are decoded, interpreted in context, fused, and then presented so that it can be understood and might take the form of a chart, a report, statistics, or some aggregation of information.

5.1.2 Diagnostic Analytics. It is the process of understanding why something has happened. This goes one step deeper than descriptive analytics in that we try to find out the root cause and explanations for the IoT data. Both descriptive and diagnostic analytics give us hindsight on what and why things have happened.

5.1.3 Discovery in Analytics. Through the application of inference, reasoning, or detecting non-trivial information from raw IoT data, we have the capability of Discovery in Analytics. Given the acute problem of volume that big data presents, Discovery in Analytics is also very valuable in narrowing down the search space of analytics applications. Discovery in Analytics on data tries to answer the question of what happened that we do not know about, and the outcome is insight into what happened. What differentiates this from the previous types of analytics is using the data to detect something new, novel, or different (e.g., trends, exceptions, or clusters) rather than describing or explaining it.

5.1.4 Predictive Analytics. For the final two categories of analytics, we move from hindsight and insight to foresight. Predictive Analytics tries to answer the question, "What is likely to happen?" It uses past data and knowledge to predict future outcomes [76] and provides methods to assess the quality of these predictions [184].

5.1.5 Prescriptive Analytics. It looks at the question of what should I do about what has happened or is likely to happen. It enables decision-makers to not only look into the future about opportunities (and issues) that are potentially out there, but it also presents the best course of action to act on foresight in a timely manner [22] with the consideration of uncertainty. This form of analytical capability is closely coupled with optimisation, answering "what if" questions to evaluate and present the best solution.

5.2 Specific Types of Analytics

Having looked at analytical capabilities, which help to define the aims of analytics, we look at specific analytics that can guide stakeholders involved in the deployment of analytics on IoT applications. A summary of the specific types of analytics and their corresponding analytical capabilities can be found in Figure 5.

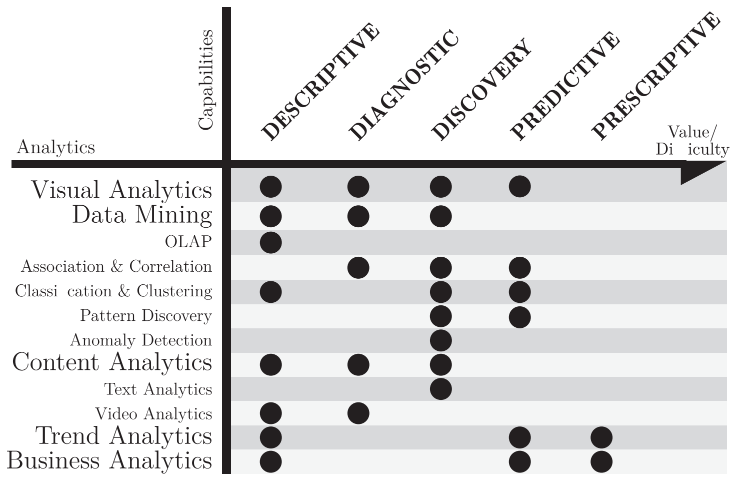

Fig. 5. Classification of types of analytics.

5.2.1 Visual Analytics. Visual analytics combines interactive visualisations with data analytics techniques "for an effective understanding, reasoning, and decision making on the basis of very large and complex data sets" [106]. Hence, visual analytics can contribute to not only describing and diagnosing what happened but also help users to discover new insights. In the work by Zhang et al. [219], we see visual analytics being applied to health care data and describing, through answering of questions like, "What is the distribution of pregnancy age?", diagnosing, through hypothesising two disease patterns due to "diarrhoea" and "fever" not being correlated and discovery, through detecting the delayed outbreak of two diseases.

5.2.2 Data Mining. Data Mining is part of the Knowledge Discovery from Data (KDD) process in which interesting patterns and knowledge are discovered from large amounts of data [79]. The IoT is a source for a large amount of data in which the techniques of data mining can be applied. These include:

Multi-dimensional data summary is often associated with Online analytical processing (OLAP) operations that make use of background knowledge of the domain to allow presentation of data at different levels of abstraction. For example, you could drill-down and roll-up data to present it at different degrees of summarisation.

Association & correlation is the process of finding the relationship between two variables that vary according to some pattern. This could allow us to find out whether buying product A led to buying product B with a degree of confidence and support.

Classification is the process of finding some model or function that has the ability to distinguish between data classes or concepts.

Clustering is the process of grouping data objects into classes without labels. The clustered data objects have maximum similarity to in-class objects and minimum similarity between objects from other classes.

Pattern discovery is the process of detecting and extracting interesting patterns from data, an example of which are frequent item sets, a set of items that often appear together in a transactional data set. *Anomaly detection* refers to the problem of "finding patterns in data that do not conform to expected behaviour" [34].

5.2.3 Content and Text Analytics. Content Analytics is the broad area of which analytical techniques are applied to digital content. Text analytics is the derivation of high-quality information from unstructured text, for example, extracting named entities and relations, analysing sentiment, extracting events and time-series information, and so on.

5.2.4 Video Analytics. Video Analytics (VA) is about the use of specialised software and hardware "to analyse captured video and automatically identify specific objects, events, behaviour, or attitudes in video footage in real time" [66].

5.2.5 Trend Analytics. Trend analytics is concerned with looking at data and events across time, understanding it and making predictions to future trends and providing early warning systems. Trend analytics is also closely related to the analysis of time-series information [35], where looking at a time series we try to find a "long-term change in the mean level."

5.2.6 Business Analytics. Business Analytics is the practice of using an organisations data to gain insights through analytical techniques that can better inform business decisions and automate and optimise business processes.

5.3 A Layered Taxonomy of Data, Analytics and Applications for the IoT

Figure 6 shows a layered taxonomy of analytics for the IoT that summarises our survey with respect to analytics capabilities and specific analytics. There are three layers in the taxonomy: data, analytics, and applications. Within each layer are various concepts, classes, and techniques that are well-defined in background literature and gathered from reviews in each area.

In the analytics layer, visual analytics processes are defined by Keim et al. [107], while techniques for each data type are summarised in surveys [5, 134, 189]. Data mining [67, 114, 183], text analytics [3], and video analytics [117] each are well-described in the referenced authoritative texts. Time-series forecasting [123], analysis, and control [29] have also been reviewed in detail. Literature also covers business analytics processes [111], prescriptive analytics [22], and techniques [193].

In the application layer, themes and domains are from Section 3.2, while the IoT applications from each domain surveyed in Section 4 are shown connected to their various analytics capabilities. Analytics techniques can then be referenced under each capability.

In the data layer, big data as defined in Section 1 is summarised along with terms used throughout the survey, including currency, types of data, and their sources. Two other terms for big data, veracity, and variability are introduced for completeness. Veracity is concerned with the noise within data and how accurate the data is for whatever purpose it is to serve. Variability is concerned with data whose meaning changes due to differences in interpretation of data within a specific context. Finally, processes, distribution levels, and distributed technologies for storage and compute are covered in Section 6.

6 ENABLING INFRASTRUCTURE FOR IOT ANALYTICS

In the previous section, we looked at classifying analytics and building a taxonomy for understanding analytics. In this section, we will review work that enables analytics to be applied on IoT data.

Enabling infrastructure for analytics on the IoT are components, techniques, and technology that contribute to the process whereby data is utilised in analytics applications. Figure 7 shows the process of how data goes through the steps of generation and collection, aggregation, and integration and finally is applied in analytics applications [38]. Storage and compute are abstract processes involved with each step of this data flow. In practice, data could be pipelined from one

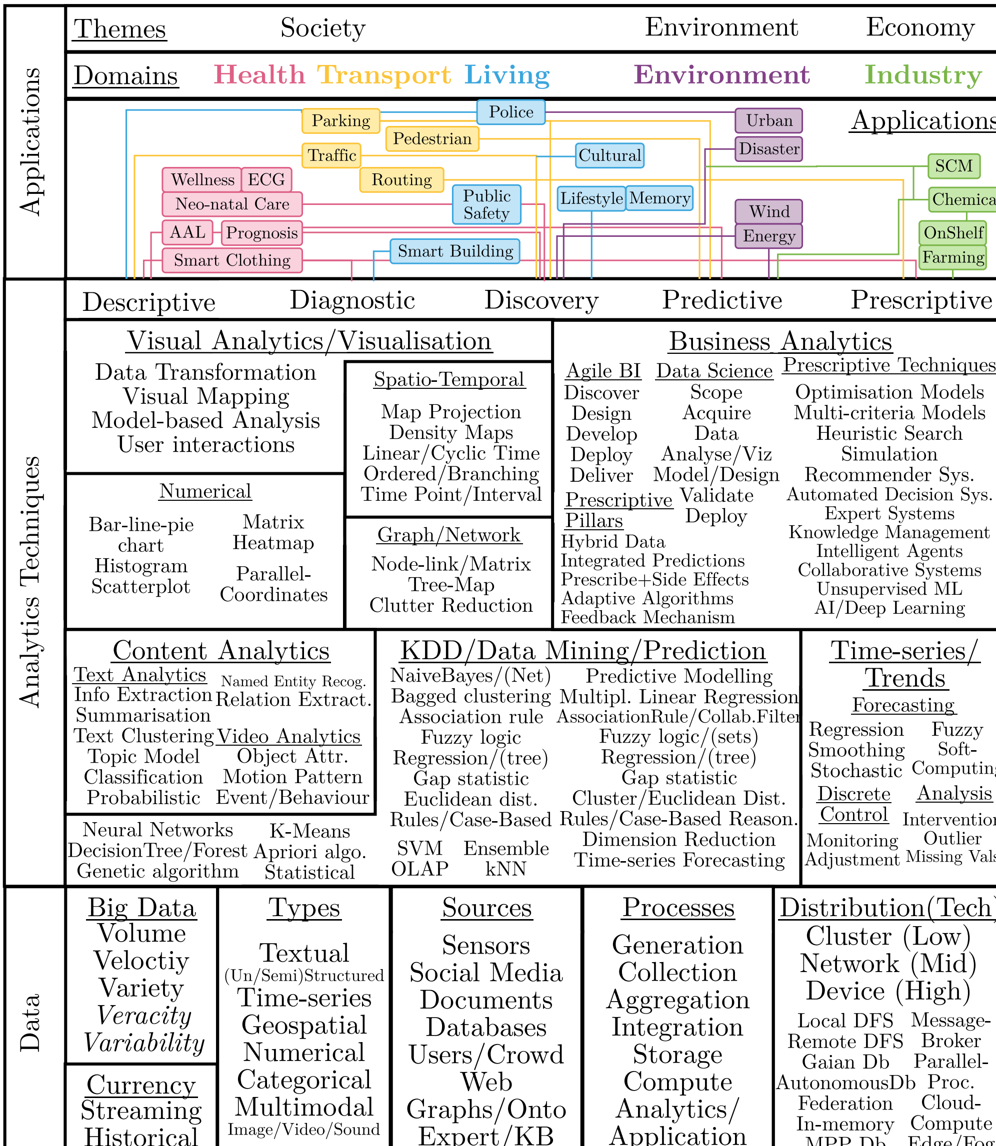

Fig. 6. Layered taxonomy of analytics from data to application.

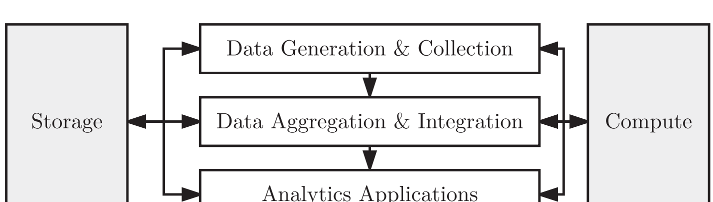

Fig. 7. Data flow process for analytics applications.

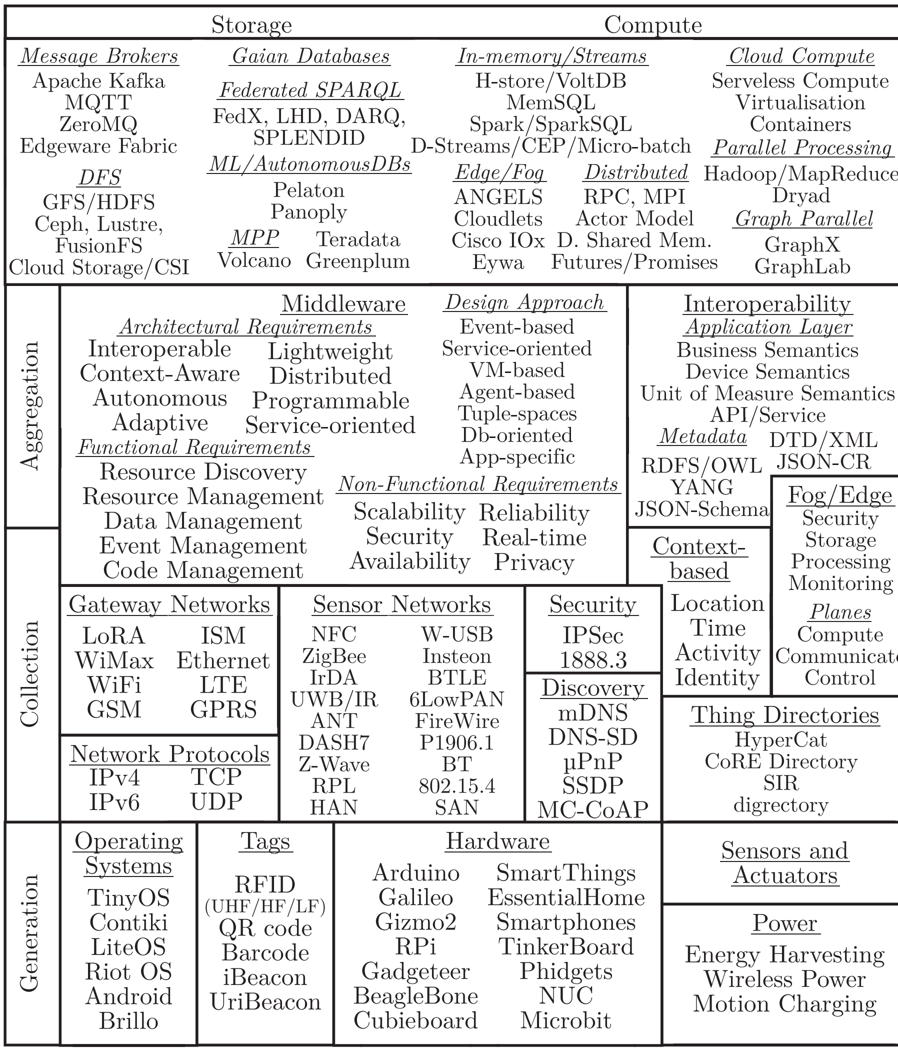

Fig. 8. IoT enabling infrastructure for analytics: generation, collection, aggregation, storage, compute.

step to another, hence, need not necessarily be stored, physically, in a separate location. Compute could also be done on the device or in transit and need not imply a separate compute component.

The following sections elaborate on each step of the data flow in IoT analytics from data generation and collection to aggregation and integration with storage and compute alongside. Figure 8 summarises the technologies covered.

6.1 Data Generation: Sensors and Tags, Hardware and OS, Power

A major source of data in the IoT is generated from sensors including many types of environmental, spatial sensors and health sensors [165]. Tags also generate data and can be passive like QR code and barcode patterns, which require a device to scan or be active like iBeacon [13] and UriBeacon [69] technologies, which project signals to mobile applications. RFID tags can be either passive or active, the active type requiring a power source to broadcast signals and can be UHF (Ultra High

Frequency), HF (High Frequency), or LF (Low Frequency). A list of hardware platforms for sensors or base stations receiving the generated data and a list of lightweight operating systems for the IoT are discussed in the surveys by Ray [166] and Razzaque et al. [168], respectively.

Remotely deployed IoT sensors also require power especially for the energy consuming process of wirelessly transmitting data. Wolf [205] describes a number of energy scavenging systems that harvest energy from the environment, while wireless charging technologies like ubeam [194] and motion charging like Ampy [12] are alternatives. Data is then transmitted and collected as follows.

6.2 Data Collection: Discovery, Management, Transmission, Context, and Fog

A significant amount of work on the IoT has been to develop middleware, the software layer that connects various components like the device, storage, compute and network together. Middleware in the IoT has functional requirements [19, 168] including: (1) resource discovery, (2) resource management, (3) data management, (4) event management, and (5) code management. Of these requirements, resource discovery and management fit within the collection step while data and event management fit within the aggregation and storage processes while code management fits within the compute process.

There are a number of technologies for the IoT that support the discovery of devices, Multicast DNS (mDNS) [94], DNS Service Discovery (DNS-SD) [40], Micro Plug and Play (μPnP) [211], Simple Service Discovery Protocol (SSDP) [92], and Multicast CoAP (MC-CoAP) [95]. One means of managing the discovered resources is through Thing Directories that serve as catalogues of resources. HyperCat [8], CoRE Resource Directory [96], Sensor Instance Registry (SIR) [101], and digrectory [98] are various implementations supporting resource lookup and search.

Another important process in data collection is the transmission of generated data. We divide the transmission technologies into those for communication within sensor networks like Zigbee and those for communication within gateway networks and the wider IoT like LTE and GSM. These technologies are discussed in the survey by Ray [165] on IoT architecures. Network and transport layer protocols like IPv4/v6 and TCP/UDP are well-defined in literature. IPSec [93] is a security protocol suite for the network layer that authenticates and encrypts packet data, while 1888.3 [90] is a security standard for the IEEE Ubiquitous Green Community Control Network.

Context-based computing is a research area within the IoT that involves the detection, sharing and grouping of devices according to context in the IoT. Context from the conceptual perspective, as described by Perera et al. [151], refers to the location, time, activity, and identity related to data collected. Grim et al. [73] design a bloom filter [28] inspired data structure that summarises this context and identifies set membership in a probabilistic way so resources can be discovered and grouped. Perera et al. [150] implement resource search and management on a context-based framework. A Comparative Priority-based Weighted Index is generated for each resource, combining priorities like accuracy, reliability, energy, cost, and availability, which optimises the selection process for the aggregation of data sources.

Chiang et al. [43] define fog computing as an “end-to-end horizontal architecture” for the IoT that distributes the compute and storage, control and communication planes nearer to users “along the cloud-to-thing continuum.” Aazam and Huh [1] describe specifically how this vision can be realised in terms of additional security, storage, processing and monitoring sub-layers between the physical layer and the transport layer of an IoT architecture. Hence, Fog Computing extends to the data aggregation layer and can even extend to the analytics process.

6.3 Data Aggregation and Integration: Interoperability

Besides the functional requirements of middleware defined in the previous section, the survey by Razzaque et al. [168] also describes architectural requirements, design approaches, and non-functional requirements of middleware, as shown in Figure 8, along with a detailed review of various software and publications. Interoperability is one of the architectural requirements and is essential for the data aggregation and integration process. McKinsey [125] estimate that such interoperability will unlock an additional 40 to 60 percent of the total projected future IoT market value.

Berrios et al. [25] describe how various cross-industry consortia concerned with the IoT are converging on semantic interoperability within the application layer, which they split into interoperability of business semantics, device semantics, unit of measure semantics and API and service standards. All the consortia involved are working on device semantics for interoperability, while at least one consortium has defined standards for each of the home & buildings, retail, healthcare, transport & logistics, and energy industries. The series of articles, co-authored by representatives from each consortia, also recommended a top-level ontology, an ontology representing the intersection of business and device semantics and a common data format.

Milenkovic [132] also argue for a common representation for metadata that provides context to the data collected. Linked Data, which is defined as “a set of best practices for publishing data on the Web so that distributed structured data can be interconnected and made more useful by semantic queries” [27], is seen as one means. Barnaghi et al. [21] argue that Linked Data and semantic technologies can serve to facilitate interoperability, data abstraction, access and integration with other cyber, social or physical world data. RDFS, which inspired the popular schema.org vocabulary that allows persons, events, places, and products to be defined on the web and the Web Ontology Language (OWL) for complex modelling and non-trivial automated reasoning in ontologies are related technologies that allow metadata to be represented. There are also proposals for other data models like YANG [179], JSON Schema [62], and JSON Content Rules [46] to be adopted.

6.4 Architectures for Storage and Compute

At a high level, architectures help to define how to build infrastructures and how to handle big IoT data in the storage and compute components for analytics. One such architecture is the lambda architecture by Marz et al. [128], which consists of a speed, a serving, and a batch layer. The idea is that for huge datasets it is necessary to precompute batch views in the batch layer and update them in the serving layer, at the same time a speed layer compensates for the high latency of the batch computations by looking at recent data and doing fast incremental updates.

This big data architecture is useful in providing us with a general idea of how analytics can scale to the volume of IoT data. Ye et al. [212] implement a service for big data analytics (using R and Hadoop for efficient parallel processing [48]) in the batch layer to do data-mining tasks like clustering. Products like Onix [185], which do analytics on streams, work on implementing solutions for the speed layer while industry players like MapR [127] have also proposed the Lambda Architecture as part of their data-processing architecture. The Lambda Architecture has also been used in an IoT context by Villari et al. [201], who apply it to a Smart Environment use case.

Baldominos et al. [18] also propose a design that is similar in structure to the Lambda architecture and is another example of how an analytics system, for doing machine learning and recommendations in this case, can be implemented with this separation of batch (batch machine-learning module/storage), speed (stream machine-learning module), and serving (dashboard) layers.

The Hadoop and Spark ecosystems are two other big data-processing architectures. Hadoop consists of two main parts, a Distributed File System (DFS) like HDFS, and a distributed programming model like MapReduce. The Hadoop ecosystem[1] consists of various technologies built on top and around these two parts including warehousing like Hive, NoSQL databases like HBase,

[1]Available from http://thebigdatablog.weebly.com/blog/the-hadoop-ecosystem-overview.

data ingestion pipes like Flume, and machine-learning libraries like Mahout and a host of other technologies.[2]

The Spark ecosystem is built on Spark Core and consists of components like SparkSQL, Spark Streaming, MLLIB and GraphX amongst others. Spark is described in more detail in Section 6.6.

The Lambda architecture, Hadoop, and Spark ecosystems, however, are suited for big data systems in which compute and storage are in centralised or cloud-based clusters rather than decentralised fog and edge based computing. The next two sections describe storage and compute technologies that can be used for the IoT and big data analytics including fog computing technologies. Table 4 summarises the distributed storage and compute technologies and their references.

6.5 Storage Technologies

Storage file systems need to cope with the huge amount of data from the IoT and work on "exascale" file systems by Raicu et al. [161] look to address issues of scalability to millions of nodes and billions of concurrent input/output requests. The idea is to combine advances in non-volatile storage with those of distributed file systems. These include the management of distributed metadata, partitioning and knowledge of data access patterns to maximise data locality, resilience and high availability, data indexing and cooperative caching. An implementation exists in the form of FusionFS [162], which implements a zero-hop distributed hash-table (ZHT) for metadata management.

Similar decentralised distributed file systems (DFS) like Ceph [204] and GlusterFS also manage metadata in a distributed way, while other DFS like HDFS, which is part of the Hadoop ecosystem from Section 6.4, iRODS and Lustre [121] are centralised with a single or replicated metadata servers. This group of DFS are classified as locally managed DFS and are compared in a survey by Depardon et al. [53]. Another group of DFS are remote access DFS like cloud storage from Google Cloud Storage [68], S3 [10], and Azure Blob [131]. Another interesting dimension to remote access DFS is the emerging Container Storage Interface (CSI) specification [83] for provisioning and managing storage, including cloud DFS like Quobyte [160], from container applications.

Bent et al. [23] have designed a distributed, federated database architecture, Gaian Databases, that uses biologically inspired, self-organising principles to organise a network of heterogenous relational or flat file databases and enable queries across them through query flooding. The work has become part of IBM's Smarter Planet [89] project—an IoT-related vision of the planet that together with Edgware Fabric [88] form a middleware layer for analytics and intelligence. The advantage of Gaian Databases is that through minimising network diameter and maximising connections to fit nodes, analytical queries on distributed data can be performed quickly and reliably.

Linked Data was seen as an approach to the aggregation step previously and work to access Linked Data across distributed sources has led to the area of federated querying. FedX [180], SPLENDID [70], LHD [203], and DARQ [159] are all engines that optimise federated query performance. They achieve improved performance by optimising the join order in queries. FedX takes a heuristic approach while the other engines take statistical approaches. Saleem et al. [174] and Hartig [81] review and evaluate the systems. More specific performance bottlenecks like data distribution [163] and other challenges [164] for federated engines have still to be addressed though.

A message broker is an intermediary that routes a message from publishers to subscribers. A message broker can serve as a storage and interoperability technology in distributed systems as it can provide a formal message protocol for publishing and subscribing, reliable storage and guaranteed message delivery. Log-structured storage has been utilised for high throughput distributed message brokers like Kafka [100] or the scalable data middleware for smart grids described by

[2] Available from https://hadoopecosystemtable.github.io/.

Table 4. Summary of Distributed Storage and Compute Technologies

Technology	Product	Remark
Locally managed DFS	GFS/HDFS [63]	Centralised Block Storage
	Lustre [121]	Centralised Object Storage
	Ceph [204]	Decentralised Object Storage
	FusionFS [162]	ZHT, Data Access Paritions
Remote access DFS	Cloud Storage [10, 68]	Google Cloud Storage, S3, Azure Blob
	Container Storage [83]	DFS For Containers [160]
Message Brokers	Apache Kafka [100]	Log Structured Broker
	MQTT [118]	Lightweight Pub/Sub Protocol
	ZeroMQ [84]	Lightweight Messaging Library
	Edgware Fabric [88]	IoT Service Bus (Discovery+Routing)
Gaian Databases	GaianDb [23]	Self-organising Network
Autonomous/ML DB	Pelaton [148]	Classify, Forecast, Optimise Workload
	Panoply [146]	ML Self-Optimisation
Federated SPARQL	FedX [180]	Optimise Join Order (Heuristic)
	SPLENDID [70]	Optimise Join Order (Statistical)
	LHD [203], DARQ [159]	Optimise Join Order (Statistical)
In-memory	H-Store [102]	Partitions, Stored Procedures
	MemSQL [129]	Real-time Data Warehouse
	Spark [215]	In-memory DAG execution
MPP/Parallel Databases	Greenplum [156]	Master, Segment PostgresSQL
	Teradata Database [190]	Shared Nothing OLTP/OLAP
	Volcano [71]	Exchange Meta-operator
Data Parallel	Hadoop/MapReduce [52]	Big Data Programming Model
	Dryad [97]	Data Parallel App Runtime
Graph Parallel	GraphX [207]	Resilient Distributed Graph Transform
	GraphLab [119]	Asynchronous, Dynamic Computation
Cloud Compute	Virtualisation	EC2 [10], Compute Engine [68]
	Serverless	Lambda [10], Functions [68, 131]
	Container [31]	Portability, Overhead, Orchestration
Edge/Fog Compute	ANGELS [137]	Partition Data, Schedule Fog Jobs
	Eywa [187]	Distributed Stream Processing
	Cloudlets [177]	Proximity, Virtualisation
	Cisco IOx [44]	Fog Director, App Host/Manage

Legend: DB=Database, DFS=Distributed File System, ML=Machine Learning, MPP=Massively Parallel Processing, OLAP/OLTP=Online Analytical/Transaction Processing, ZHT=Zero-hop Distributed Hash Table.

Yin et al. [214]. MQTT [118], a lightweight publish-subscribe protocol for the IoT, ZeroMQ [84], a messaging protocol library and Edgware Fabric [88], an IoT service bus, are other examples of technologies used in distributed message broker systems.

Autonomous or self-driving database management systems like Pelaton [148] integrate artificial intelligence components to automatically classify and forecast workloads so that the database can optimise physical storage, data location, and partitioning in distributed or cloud-based environments and runtime resources, configuration, and query cost models. Panoply [146] is a similar machine-learning optimised autonomous data warehouse, which additionally allows the "self-preparation" (automated transformation and integration) of ingested semi-structured data.

Massive Parallel Processing (MPP) databases are built on top of shared-nothing MPP grids where data is shared between nodes and nodes processes computations, queries to retrieve and process data, in parallel. Greenplum [156], which uses a master-segment approach with each segment a PostgresSQL database, and Teradata [190] are examples of MPP databases. Volcano [71] was an early system that presented research on parallelising query operators through the exchange of meta-operators.

6.6 Compute Technologies and IoT Analytics Applications

The elasticity of resources on the cloud is often considered an advantage for deploying horizontally scalable parallel processing paradigms that work with big IoT data. Compute on the cloud can be divided into virtualisation, serverless computing, and container technologies and orchestration. Major vendors like the Google Cloud Platform [68], Amazon Web Services [10], and Microsoft Azure [131] each have options for virtualisation, Compute Engine, EC2, and Azure Virtual Machines, respectively, which allow a full server to be provisioned for compute and storage tasks. Each also supports the serverless execution of compute functions through Cloud Functions, Lambda, and Azure Functions, respectively. Finally, container technologies [31] are becoming increasingly popular as they increase application portability and reduce dependencies, have lower overhead and faster launch times than virtual machines, and the orchestration of containers allows the efficient provisioning, deployment, and management of distributed compute clusters. Kubernetes, Docker Compose [191], and Mesosphere [130] are such container orchestration technologies.

Various IoT infrastructures and deployments have implemented distributed cloud-based compute. Xu et al. [209] have developed a cloud-based time-series analytics platform for the IoT that stores and indexes time-series data, analyses, and mines for patterns and allows searching on patterns and abnormal pattern discovery. Indexes specifically optimised for time-series data help achieve real-time analytics at a lower latency (at the cost of increased storage space). Ding et al. [55] propose a means of doing statistical analysis on the cloud. Spatial aggregation of the area in a city where the pollution level is above a certain threshold or parameter aggregation to calculate the average pollution level at a certain time in a city are examples. The novel part of this approach is that analytics is implemented within the database kernel itself, improving performance by reducing the transfer of data (to the master node for processing).

Nastic et al. [139] have designed a high level programming model abstraction for the IoT running on the cloud. In the model, there are abstractions called Intents and Intent Scopes, which describe a task and a group of "things," respectively, from underlying distributed and heterogenous sources that share a common context. By coupling Intents and Intent Scopes with analytics operators, complex IoT applications can be designed, optimised on a distributed compute system, and run on the cloud. Guazzelli et al. [74] make use of the Predictive Model Markup Language (PMML) [49], an XML-based markup language to describe data-mining models, to run analytics on the cloud. Web service calls can be made to instances on the cloud, submitting markup that then execute tasks like running regression models, clustering, learning based on artificial neural networks (ANN), decision trees, support vector machines, or mining association rules.

Next, we briefly summarise specific distributed compute technologies from the small programming constructs and components used to build distributed compute to the large big data systems made from these components, which we divide into in-memory and stream systems, parallel programming models, graph parallel models, and edge/fog computing systems.

A means that nodes in a distributed compute system can communicate is through message passing. A Remote Procedure Call (RPC) is a form of message passing and gRPC, a multiplexed, bi-directional streaming RPC protocol, and Thrift [158], an asynchronous RPC system, are examples. The Actor Model [77] is a message passing programming model supporting asynchronous communication in distributed compute systems the provides an abstraction enabling looser coupling among components, allows for behaviour reasoning, and provides a lightweight concurrency primitive across machines. Futures or Promises are another construct for asynchronous programming and are abstractions of values that will eventually become available. They are useful for message passing in distributed compute to reason about state changes when latency is a concern.

Distributed in-memory databases like H-store [102] and MemSQL [129] allow low latency stored procedures and interactive querying, respectively, to be done on scale-out transactional databases, hence, overcoming memory limitations by adding nodes. They are so fast that they can be used for real-time compute tasks rather than just storage. Spark [215] is an in-memory data-processing engine with two main abstractions, an immutable, read-only collection of objects within Resilient Distributed Datasets (RDD) (as opposed to fine-grained Distributed Shared memory (DSM)) and parallel operations represented as an acyclic data flow graph. SparkSQL includes an execution model that uses the Catalyst query optimiser [14] for both rule-based and cost-based optimisation to form a Spark data flow graph. D-streams is the Spark streaming abstraction where a streaming computation is treated as series of deterministic batch computations on RDDs within small time intervals. This type of stream processing is called micro-batch processing while Complex Event Processing (CEP) [120] involves continuous operators on each tuple. Khare et al. [109] show how continuous operators can work on publish-subscribe IoT sensor streams with a Functional Reactive Programming (FRP) language. The system was tested on sensor data of a football match to aggregate running data for each player and create descriptive analytics heat maps for players.

Data parallelism means that each node in a distributed compute system can perform independent calculations on a meaningful subset of data. MapReduce [52], of which Hadoop MapReduce (Section 6.4) is an implementation, and Dryad [97] are both programming models for data parallel processing on big data. Hammond et al. [78] deploy analytics in the cloud using Hadoop MapReduce. The analytics techniques include text classification using Naive Bayes, a top-K recommendation engine based on similarity and a Random Forests classifier to categorise data as part of Decision Support Systems. MapReduce, however, does not scale easily for iterative graph algorithms as each iteration requires reading and writing results to disk. Graph Parallel abstractions like those in GraphX [207], for graph transformations, and GraphLab [119], for asynchronous computation, support these.

Finally, Fog or Edge Computing technologies like ANGELS [137] and the scheduler designed by Dey et al. [54] propose utilising the idle computing resources of edge devices like smartphones through a scheduler in cloud. The edge devices themselves keep track of their resource usage states, which are formed based on user behavioural patterns, and advertise free slot availability. The cloud servers receive analytics jobs and advertisements from edge devices and then schedule subtasks to these devices. Distributed stream processing within a fog computing network has also been implemented in the Eywa framework [187] using inverse-publish-subscribe for the control plane and workload pushdown to fog nodes for projections in the data plane. Cloudlets [177] allow a mobile user to instantiate virtualised compute tasks on physically proximate cloudlet hardware. Cisco IOx [44] is another platform that consists of a fog director, application host, and management components, allowing fog computing tasks to be virtualised and executed on fog nodes.

6.7 Levels of Distribution of Storage and Compute

The Internet of Things, as defined in Section 3, comprises smart and interconnected physical objects with varying storage and compute capabilities. Analytics processing can be done at different distribution levels depending on how far data from physical objects can and should travel and on the storage and compute capabilities at each of:

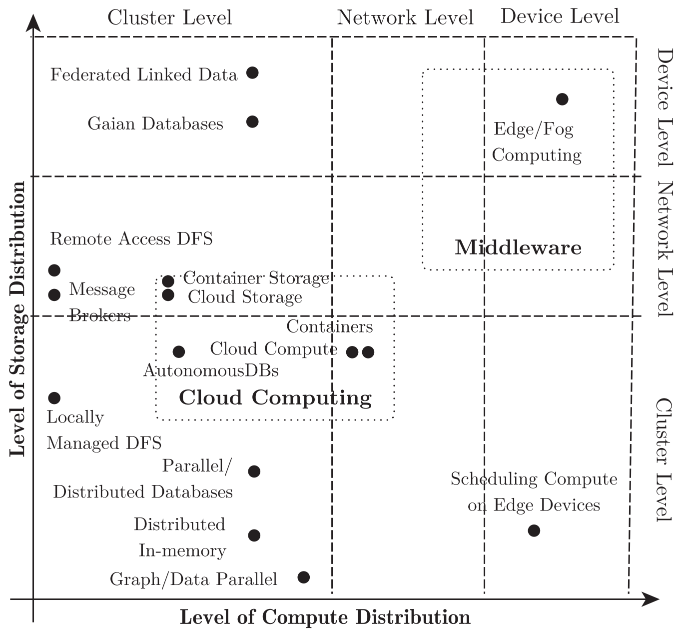

Fig. 9. Technology and their levels of distribution for storage and compute.

(1) the device level, where devices act not just as data producers but as participants of the storage and compute process,
(2) the network level, involving remote connections to fog computing nodes, hubs, base stations, gateways, routers, and servers, and
(3) the cluster level, within a group of interconnected servers.

Enabling infrastructure and technologies are observed to address each of these levels of distribution of compute and storage to a different degree. A classification of the surveyed IoT enabling technologies is proposed in Figure 9 along the axes of storage and compute distribution.

At the cluster level, we see storage systems that are distributed within locally managed clusters. Usually these clusters are located within data centres and connected by top-of-the-rack switches in a hierarchical fashion (intra and inter rack). Locally managed distributed file systems are an example. In-memory systems distribute both processing and storage, usually by partitioning the data onto nodes, in a centrally managed cluster and running processing on each node that corresponds with the data on that node. Similarly, Massive Parallel Processing Databases, Data Warehouses, and Parallel/Distributed Databases are examples of systems with distributed storage and compute on each node. Examples of compute within clusters of distributed servers include Parallel Processing frameworks like MapReduce [52].

Cloud Computing is usually divided into private, public or hybrid clouds. Private clouds share similarities and types of distributed storage and compute with those previously mentioned at cluster level. Public clouds are remotely managed and hence belong to the network level of distribution of which Cloud Storage and Cloud Compute Engines serve storage and compute tasks, respectively. Hybrid clouds bridge both public and private clouds. Similar to cloud storage are remote access distributed file systems. Message brokers, message queues, and log-based systems are some other examples of network level, possibly remote access storage systems.

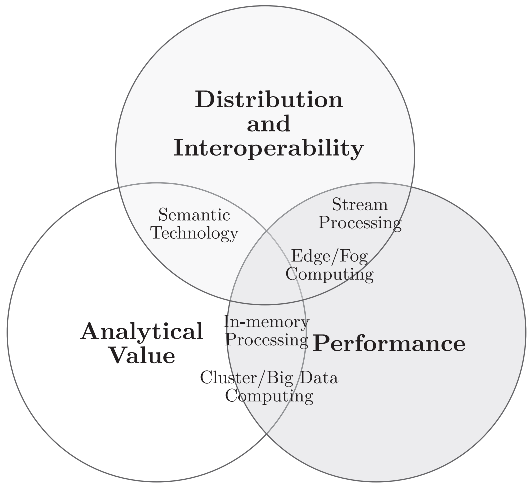

Fig. 10. Tradeoffs in designing for analytics on the IoT.

At the device level of storage, we have technologies like federated Linked Data endpoints and Gaian databases where data can reside on their respective devices but be accessed by other clients. On the device level for compute, scheduling of compute tasks on fog or edge devices, is an example. Finally, both compute and storage distribution from the network to device level is present in Edge and Fog computing and middleware is usually used to connect such edge systems together.

Table 4 summarises the surveyed literature on distributed storage and compute technologies to provide a point of reference for researchers on current state-of-the-art implementations.

This review of enabling infrastructure and technologies at each part of the data flow process, classification of the storage and compute distribution and examples of distributed storage and compute technologies form a basis for a direction of future work towards tackling the challenges big data analytics on the IoT.

7 RESEARCH CHALLENGES

As we have seen in our study of enabling infrastructure and present analytical applications in the IoT, there are still some challenges that we face in aligning the vision of the IoT with that of analytics. In particular, we argue that infrastructure for analytics in the IoT faces a tradeoff between:

(1) Distribution & Interoperability, complicated by big data *variety*,
(2) Performance, complicated by the *volume and velocity* of the big data problem,
(3) and Analytical Value, which deals with how high the output of analytics applications is on the knowledge hierarchy from Figure 4.

Figure 10 depicts the tradeoffs that IoT infrastructure for analytics applications face in terms of these three challenges. For example, the semantic technology community argues for its utility in the IoT [21] to encourage semantic interoperability, while semantic ontologies provide analytical value and federation supports diverse, heterogeneous distributed sources. Performance of such systems, though, are still questionable [164, 174]. Edge and fog computing is also an emerging area of distributed technologies that promises advantages in latency for real-time processing of streams and efficiency due to its proximity sources [43]. However, cloud-based clusters and fast-distributed OLTP in-memory processing still offer greater analytical value, combining big data sans the advantages of IoT distribution and interoperability.

Variety has been a less researched aspect of the big data problem but is apparent in the IoT paradigm. Heterogenous data sources in the IoT combined with the need for analytics to also involve a wide range of multi-modal data sources like social media, Linked Data, image and video data, satellite and geospatial data, voice data, and so on, makes the variety problem highly analogous with the richness of insights and knowledge that can be derived in analytics applications. Predictive analytics can be made more accurate through corroboration of independent data sources and prescription can be optimised with more and varying knowledge inputs. Solving the variety problem can be seen as an opportunity to enhance the value of current IoT applications.

Performance and scalability questions still exist in current systems because of the scale of the IoT. This is not only about scaling the storage of data or of the communications layer but also the scaling of infrastructure to do analytics processing. We see distributed analytics as a plausible means of handling IoT scale-data (which is predicted to be larger and richer than web scale data) and there is potential for more work in this area.

8 CONCLUSIONS

The Internet of Things (IoT) has huge potential to provide advanced services and applications across many domains, and the momentum that it has generated, together with its broad visions, make it an ideal frontier for pushing technological innovation. We have shown that analytics plays a role in many applications, across many domains, designed for the IoT and will be even more important in the future as the enabling infrastructure develops and scales and the deployment of devices becomes truly ubiquitous. We have applied a systematic review of analytics applications in the IoT to the task of understanding analytics as it develops. This results in a layered taxonomy that defines and categorises analytics by their capabilities and application potential for research and application roadmaps. We then review the enabling infrastructure and discuss the technologies from different stages in the data flow for analytics. Finally, we look at some tradeoffs for analytics in the IoT that can shape research direction going forward.

REFERENCES

[1] Mohammad Aazam and Eui Nam Huh. 2014. Fog computing and smart gateway based communication for cloud of things. In *Proceedings of the International Conference on Future Internet of Things and Cloud.* Retrieved from DOI:https://doi.org/10.1109/FiCloud.2014.83

[2] Estefania Abad, Francisco Palacio, M. Nuin, Alberto G. Zárate, A. Juarros, José María Gómez, and Santiago Marco. 2009. RFID smart tag for traceability and cold chain monitoring of foods: Demonstration in an intercontinental fresh fish logistic chain. *J. Food Eng.* 93, 4 (2009), 394–399. Retrieved from DOI:https://doi.org/10.1016/j.jfoodeng.2009.02.004

[3] Charu C. Aggarwal and Cheng Xiang Zhai. 2012. *Mining Text Data.* Springer. Retrieved from https://dl.acm.org/citation.cfm?id=2669206.

[4] Hussnain Ahmed. 2014. *Applying Big Data Analytics for Energy Efficiency.* Masters Thesis. Aalto University. Retrieved from https://aaltodoc.aalto.fi/handle/123456789/13899.

[5] Wolfgang Aigner, Silvia Miksch, Wolfgang Müller, Heidrun Schumann, and Christian Tominski. 2008. Visual methods for analyzing time-oriented data. *IEEE Trans. Visual. Comput. Graph.* 14, 1 (2008), 47–60. Retrieved from DOI:https://doi.org/10.1109/TVCG.2007.70415

[6] Jacky Akoka, Isabelle Comyn-Wattiau, and Nabil Laoufi. 2017. Research on big data—A systematic mapping study. *Comput. Stand. Interfaces* 54, 2 (2017), 105–115. Retrieved from DOI:https://doi.org/10.1016/j.csi.2017.01.004

[7] Ala Al-Fuqaha, Mohsen Guizani, Mehdi Mohammadi, Mohammed Aledhari, and Moussa Ayyash. 2015. Internet of things: A survey on enabling technologies, protocols and applications. *IEEE Commun. Surveys Tutor.* 17, 4 (2015), 2347–2376. Retrieved from DOI:https://doi.org/10.1109/COMST.2015.2444095

[8] Hypercat Alliance. 2017. Hypercat. Retrieved from http://www.hypercat.io/.

[9] Ignacio González Alonso, María Rodríguez Fernández, Juan Jacobo Peralta, and Adolfo Cortés García. 2013. A holistic approach to energy efficiency systems through consumption management and big data analytics. *Int. J. Adv. Softw.* 6, 3 (2013), 261–271. Retrieved from http://digibuo.uniovi.es/dspace/bitstream/10651/35765/1/soft.

[10] Amazon Web Services. 2015. AWS. Retrieved from http://aws.amazon.com/products/.

[11] Sara Amendola, Rossella Lodato, Sabina Manzari, Cecilia Occhiuzzi, and Gaetano Marrocco. 2014. RFID technology for IoT-based personal healthcare in smartspaces. *IEEE Int. Things J.* PP, 2 (2014), 1–1.

[12] Ampy. 2017. Ampy live charged. Retrieved from http://www.getampy.com/.

[13] Apple. 2017. iBeacon. Retrieved from https://developer.apple.com/ibeacon/.

[14] Michael Armbrust, Ali Ghodsi, Matei Zaharia, Reynold S. Xin, Cheng Lian, Yin Huai, Davies Liu, Joseph K. Bradley, Xiangrui Meng, Tomer Kaftan, and Michael J. Franklin. 2015. Spark SQL: Relational data processing in spark. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. Retrieved from DOI: https://doi.org/10.1145/2723372.2742797

[15] Niccolo Aste, Massimiliano Manfren, and Giorgia Marenzi. 2017. Building automation and control systems and performance optimization: A framework for analysis. *Renew. Sustain. Energy Rev.* 75, 2017 (2017), 313–330. Retrieved from DOI: https://doi.org/10.1016/j.rser.2016.10.072

[16] Luigi Atzori, Antonio Iera, and Giacomo Morabito. 2010. The internet of things: A survey. *Comput. Netw.* 54, 15 (Oct. 2010), 2787–2805. Retrieved from DOI: https://doi.org/10.1016/j.comnet.2010.05.010

[17] Antoine Bagula, Lorenzo Castelli, and Marco Zennaro. 2015. On the design of smart parking networks in the smart cities: An optimal sensor placement model. *Sensors* 15, 7 (2015), 15443–67. Retrieved from DOI: https://doi.org/10.3390/s150715443

[18] Alejandro Baldominos, Esperanza Albacete, Yago Saez, and Pedro Isasi. 2014. A scalable machine learning online service for big data real-time analysis. In *Computational Intelligence in Big Data*. IEEE, 1–8.

[19] Soma Bandyopadhyay, Munmun Sengupta, Souvik Maiti, and Subhajit Dutta. 2011. Role of middleware for internet of things: A study. *Int. J. Comput. Sci. Eng. Survey* 2, 3 (2011), 94–105. Retrieved from DOI: https://doi.org/10.5121/ijcses.2011.2307

[20] Oresti Banos, Muhammad Bilal Amin, Wajahat Ali Khan, Muhammad Afzal, Maqbool Hussain, Byeong Ho Kang, and Sungyong Lee. 2016. The mining minds digital health and wellness framework. *BioMed. Eng. OnLine* 15, 1 (July 2016), 76. Retrieved from DOI: https://doi.org/10.1186/s12938-016-0179-9

[21] Payam Barnaghi, Wei Wang, Cory Henson, and Kerry Taylor. 2012. Semantics for the internet of things: Early progress and back to the future. *Int. J. Semant. Web Info. Syst.* 8, 1 (2012), 1–21. Retrieved from DOI: https://doi.org/10.4018/jswis.2012010101

[22] Atanu Basu. 2013. Five pillars of prescriptive analytics success. *Analyt. Mag.* (2013), 8–12. Retrieved from http://analytics-magazine.org/executive-edge-five-pillars-of-prescriptive-analytics-success/.

[23] Graham Bent, Patrick Dantressangle, David Vyvyan, Abbe Mowshowitz, and Valia Mitsou. 2008. A dynamic distributed federated database. In *Proceedings of the 2nd Annual Conference of the International Technology Alliance*.

[24] Jay H. Bernstein. 2011. The data-information-knowledge-wisdom hierarchy and its antithesis. *NASKO 2.1* (2011), 68–75.

[25] Victor Berrios, Richard Halter, Mark Harrison, Scott Hollenbeck, Elisa Kendall, Doug Migliori, and John Petze. 2017. Cross-industry semantic interoperability. Retrieved from http://www.embedded-computing.com/semantic-interop/cross-industry-semantic-interoperability-part-two-application-layer-standards-and-open-source-initiatives.

[26] Jeff Bertolucci. 2013. Big data analytics: Descriptive vs. predictive vs. prescriptive. Retrieved from http://goo.gl/dyNDFV.

[27] Chris Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked data—The story so far. *Int. J. Semant. Web Info. Syst.* 5 (2009), 1–22. Retrieved from https://eprints.soton.ac.uk/271285/.

[28] Burton H. Bloom. 1970. Space/time trade-offs in hash coding with allowable errors. *Commun. ACM* 13, 7 (1970), 422–426. Retrieved from DOI: https://doi.org/10.1145/362686.362692

[29] George Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. 2015. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons. Retrieved from http://eu.wiley.com/WileyCDA/WileyTitle/productCd-1118675029.html.

[30] Andrea Caragliu, Chiara Del Bo, and Peter Nijkamp. 2011. Smart cities in Europe. *J. Urban Technol.* 18, January 2015 (2011), 65–82. Retrieved from DOI: https://doi.org/10.1080/10630732.2011.601117

[31] Emiliano Casalicchio. 2017. Autonomic orchestration of containers: Problem definition and research challenges. In *Proceedings of the 10th EAI International Conference on Performance Evaluation Methodologies and Tools*. Retrieved from DOI: https://doi.org/10.4108/eai.25-10-2016.2266649

[32] Marie Chan, Daniel Estève, Christophe Escriba, and Eric Campo. 2008. A review of smart homes—Present state and future challenges. *Comput. Methods Programs Biomed.* 91 (2008), 55–81. DOI: https://doi.org/10.1016/j.cmpb.2008.02.001

[33] Neil Chandler, Bill Hostmann, Nigel Rayner, and Gareth Herschel. 2011. *Gartner's Business Analytics Framework*. Technical Report. Gartner Inc. Retrieved from http://www.gartner.com/imagesrv/summits/docs/na/business-intelligence/gartners.

[34] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection. *Comput. Surveys* 41, 3 (July 2009), 1–58. Retrieved from DOI: https://doi.org/10.1145/1541880.1541882

[35] Chris Chatfield. 2013. *The Analysis of Time Series: An Introduction*. CRC Press.

[36] Hsinchun Chen, Roger H. L. Chiang, and Veda Storey. 2012. Business intelligence and analytics: From big data to big impact. *MIS Quart.* 36, 4 (2012), 1165–1188.

[37] Min Chen, Yujun Ma, Jeungeun Song, Chin Feng Lai, and Bin Hu. 2016. Smart clothing: Connecting human with clouds and big data for sustainable health monitoring. *Mobile Netw. Appl.* 21, 5 (2016), 825–845. Retrieved from DOI: https://doi.org/10.1007/s11036-016-0745-1 arxiv:1312.4722.

[38] Min Chen, Shiwen Mao, and Yunhao Liu. 2014. Big data: A survey. *Mobile Netw. Appl.* 19 (2014), 171–209. Retrieved from DOI: https://doi.org/10.1007/s11036-013-0489-0

[39] Min Chen, Shiwen Mao, Yin Zhang, and Victor Leung. 2014. *Big Data—Related Technologies, Challenges and Future Prospects*. Springer. Retrieved from http://www.springer.com/gp/book/9783319062440.

[40] Stuart Cheshire. 2017. DNS service discovery. Retrieved from http://www.dns-sd.org/.

[41] Angelo Chianese, Fiammetta Marulli, Francesco Piccialli, Paolo Benedusi, and Jai E. Jung. 2017. An associative engines based approach supporting collaborative analytics in the internet of cultural things. *Future Gen. Comput. Syst.* 66 (2017), 187–198. Retrieved from DOI: https://doi.org/10.1016/j.future.2016.04.015

[42] Leo Chiang, Bo Lu, and Ivan Castillo. 2017. Big data analytics in chemical engineering. *Ann. Rev. Chem. Biomol. Eng.* 8, 1 (2017), 63–85. Retrieved from DOI: https://doi.org/10.1146/annurev-chembioeng-060816-101555

[43] Mung Chiang, Sangtae Ha, Chih-Lin I, Fulvio Risso, and Tao Zhang. 2017. Clarifying fog computing and networking: 10 questions and answers. *IEEE Commun. Mag.* 55, 4 (Apr. 2017), 18–20. Retrieved from DOI: https://doi.org/10.1109/MCOM.2017.7901470

[44] Cisco. 2015. IOX. Retrieved from https://developer.cisco.com/site/iox/.

[45] Michael Corcoran. 2012. *The Five Types of Analytics*. Technical Report. Information Builders. Retrieved from http://www.informationbuilders.co.uk/sites/www.informationbuilders.com/files/intl/co.uk/presentations/four.

[46] Pete Cordell and Andrew Newton. 2016. A language for rules describing JSON content. *IETF Working Internet Draft*. Retrieved from https://www.ietf.org/id/draft-newton-json-content-rules-08.txt.

[47] Jay Danner, Linda Wills, Elbert M. Ruiz, and Lee W. Lerner. 2016. Rapid precedent-aware pedestrian and car classification on constrained IoT platforms. In *Proceedings of the 14th ACM/IEEE Symposium on Embedded Systems for Real-Time Multimedia*, 29–36. Retrieved from DOI: https://doi.org/10.1145/2993452.2993562

[48] Sudipto Das, Yannis Sismanis, Kevin S. Beyer, Rainer Gemulla, Peter J. Haas, and John McPherson. 2010. Ricardo: Integrating R and Hadoop. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*. Retrieved from DOI: https://doi.org/10.1145/1807167.1807275

[49] Data Mining Group. 2015. PMML 4.2—General Structure. Retrieved from http://goo.gl/t2Xvy0.

[50] Thomas Davenport. 2006. Competing on analytics. *Harvard Bus. Rev.* 84, 1 (2006), 98–107. Retrieved from https://hbr.org/2006/01/competing-on-analytics.

[51] Thomas Davenport. 2013. Analytics 3.0. Retrieved from https://hbr.org/2013/12/analytics-30.

[52] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: Simplified data processing on large clusters. *Commun. ACM* 51, 1 (2008), 1–13.

[53] Benjamin Depardon, Gaël Le Mahec, and Cyril Séguin. 2013. *Analysis of Six Distributed File Systems*. Technical Report. HAL. Retrieved from https://hal.inria.fr/hal-00789086.

[54] Swarnava Dey, Arijit Mukherjee, Himadri Sekhar Paul, and Arpan Pal. 2013. Challenges of using edge devices in IoT computation grids. In *Proceedings of the International Conference on Parallel and Distributed Systems*. Retrieved from DOI: https://doi.org/10.1109/ICPADS.2013.101

[55] Zhiming Ding, Xu Gao, Jiajie Xu, and Hong Wu. 2013. IOT-StatisticDB: A general statistical database cluster mechanism for big data analysis in the internet of things. In *Proceedings of the 2013 IEEE International Conference on Green Computing and Communications*. Retrieved from DOI: https://doi.org/10.1109/GreenCom-iThings-CPSCom.2013.104

[56] Angelika Dohr, R. Modre-Opsrian, Mario Drobics, Dieter Hayn, and Günter Schreier. 2010. The internet of things for ambient assisted living. In *Proceedings of the 7th International Conference on Information Technology*. 804–809. Retrieved from DOI: https://doi.org/10.1109/ITNG.2010.104

[57] European Commission. 2015. Digital Agenda for Europe: The Internet of Things. Retrieved from http://goo.gl/oNhYOP.

[58] Amirhossein Farahzadia, Pooyan Shams, Javad Rezazadeh, and Reza Farahbakhsh. 2017. Middleware technologies for cloud of things—A survey. *Dig. Commun. Netw.* 3, 4 (2017), 1–13. Retrieved from DOI: https://doi.org/10.1016/j.dcan.2017.04.005 arxiv:1705.00387

[59] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. From data mining to knowledge discovery in databases. *AI Mag.* 17, 3 (1996), 37–53. Retrieved from DOI: https://doi.org/10.1609/aimag.v17i3.1230

[60] Raul Castro Fernandez, Matthias Weidlich, Peter Pietzuch, and Avigdor Gal. 2014. Grand challenge: Scalable stateful stream processing for smart grids. In *Proceedings of the 8th International Conference on Distributed Event-Based Systems.* 0–5. Retrieved from DOI: https://doi.org/10.1145/2611286.2611326

[61] Amy Ann Forni and Rob Meulen. 2016. *Gartner's 2016 Hype Cycle for Emerging Technologies.* Retrieved from https://www.gartner.com/newsroom/id/3412017.

[62] Francis Galiegue, Kris Zyp et al. 2013. JSON schema: Core definitions and terminology. *IETF Working Internet Draft.* Retrieved from http://json-schema.org/latest/json-schema-core.html.

[63] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. 2003. The google file system. *ACM SIGOPS Operat. Syst. Rev.* 37, 5 (2003), 29–43.

[64] Animikh Ghosh, Ketan A. Patil, and Sunil Kumar Vuppala. 2013. PLEMS: Plug load energy management solution for enterprises. In *Proceedings of the 27th IEEE International Conference on Advanced Information Networking and Applications.* Retrieved from DOI: https://doi.org/10.1109/AINA.2013.45

[65] Bob Giddings, Bill Hopwood, and Geoff O'Brien. 2002. Environment, economy and society: Fitting them together into sustainable development. *Sustain. Dev.* 10 (2002), 187–196. Retrieved from DOI: https://doi.org/10.1002/sd.199

[66] Roberto Gimenez, Diego Fuentes, Emilio Martin, Diego Gimenez, Judith Pertejo, Sofia Tsekeridou, Roberto Gavazzi, Mario Carabaño, and Sofia Virgos. 2012. The safety transformation in the future internet domain. *Future Internet* (2012), 190–200. Retrieved from DOI: https://doi.org/10.1007/978-3-642-30241-1_17

[67] Michael Goebel and Le Gruenwald. 1999. A survey of data mining and knowledge discovery software tools. *ACM SIGKDD Explor.* 1, 1 (1999), 20–33. Retrieved from DOI: https://doi.org/10.1145/846170.846172

[68] Google. 2015. Google Cloud platform. Retrieved from https://cloud.google.com/.

[69] Google. 2017. Eddystone beacons. Retrieved from https://developers.google.com/beacons/.

[70] Olaf Gorlitz and Steffen Staab. 2011. SPLENDID: SPARQL endpoint federation exploiting VOID descriptions. In *Proceedings of the 2nd International Workshop on Consuming Linked Data.* Retrieved from http://dl.acm.org/citation.cfm?id=2887354.

[71] Goetz Graefe. 1994. Volcano—An extensible and parallel query evaluation system. *IEEE Trans. Knowl. Data Eng.* 6 (1994), 120–135. Retrieved from DOI: https://doi.org/10.1109/69.273032

[72] Jorge Granjal, Edmundo Monteiro, and Jorge Sa Silva. 2015. Security for the internet of things: A survey of existing protocols and open research issues. *IEEE Commun. Surveys Tutor.* 17, 3 (2015), 1294–1312. Retrieved from DOI: https://doi.org/10.1109/COMST.2015.2388550

[73] Evan Grim, Chien-liang Fok, and Christine Julien. 2012. Grapevine: Efficient situational awareness in pervasive computing environments. In *Proceedings of the 2012 IEEE International Conference on Pervasive Computing and Communications Workshops.* Retrieved from http://ieeexplore.ieee.org/document/6197539/.

[74] Alex Guazzelli, Kostantinos Stathatos, and Michael Zeller. 2009. Efficient deployment of predictive analytics through open standards and cloud computing. *ACM SIGKDD Explor. Newslett.* 11, 1 (2009), 32. Retrieved from DOI: https://doi.org/10.1145/1656274.1656281

[75] Bin Guo, Daqing Zhang, and Zhu Wang. 2011. Living with internet of things: The emergence of embedded intelligence. In *Proceedings of the 2011 IEEE International Conferences on Internet of Things and Cyber, Physical and Social Computing.* Retrieved from DOI: https://doi.org/10.1109/iThings/CPSCom.2011.11

[76] Joe F. Hair Jr. 2007. Knowledge creation in marketing: The role of predictive analytics. *Eur. Bus. Rev.* 19 (2007), 303–315. Retrieved from DOI: https://doi.org/10.1108/09555340710760134

[77] Philipp Haller. 2012. On the integration of the actor model in mainstream technologies. In *Proceedings of the 2nd Edition on Programming Systems, Languages and Applications Based on Actors, Agents, and Decentralized Control Abstractions.* ACM Press, New York, New York. Retrieved from DOI: https://doi.org/10.1145/2414639.2414641

[78] Klavdiya Hammond and Aparna S. Varde. 2013. Cloud based predictive analytics text classification, recommender systems and decision support. In *Proceedings of the 13th IEEE International Conference on Data Mining Workshops.* Retrieved from DOI: https://doi.org/10.1109/ICDMW.2013.95

[79] Manhyung Han, La The Vinh, Young-Koo Lee, and Sungyoung Lee. 2012. Comprehensive context recognizer based on multimodal sensors in a smartphone. *Sensors* 12, 9 (2012), 12588–12605. Retrieved from DOI: https://doi.org/10.3390/s120912588

[80] Peter E. Hart, Nils J. Nilsson, and Betram Raphael. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans. Syst. Sci. Cybernet.* 4, 2 (1968), 100–107. Retrieved from http://ieeexplore.ieee.org/document/4082128/.

[81] Olaf Hartig. 2013. An overview on execution strategies for linked data queries. *Datenbank-Spektrum* 13, 2 (2013), 89–99. Retrieved from DOI: https://doi.org/10.1007/s13222-013-0122-1

[82] Wu He, Gongjun Yan, and Li Da Xu. 2014. Developing vehicular data cloud services in the IoT environment. *IEEE Trans. Industr. Info.* 10, 2 (2014), 1587–1595. Retrieved from DOI: https://doi.org/10.1109/TII.2014.2299233

[83] Benjamin Hindman. 2017. CSI: Towards A More Universal Storage Interface For Containers. Retrieved from https://mesosphere.com/blog/csi-towards-universal-storage-interface-for-containers/.

[84] Pieter Hintjens. 2013. *ZeroMQ: Messaging for Many Applications.* O'Reilly.

[85] Jan Holler, Vlasios Tsiatsis, Catherine Mulligan, Stefan Avesand, Stamatis Karnouskos, and David Boyle. 2014. *From Machine-to-Machine to the Internet of Things: Introduction to a New Age.* Academic Press. Retrieved from https://doi.org/10.1016/B978-0-12-407684-6.00014-0.

[86] M. Shamim Hossain and Ghulam Muhammad. 2015. Cloud-assisted industrial internet of things (IIoT)—Enabled framework for health monitoring. *Comput. Netw.* 101 (2015), 192–202. Retrieved from DOI: https://doi.org/10.1016/j.comnet.2016.01.009

[87] Myriam Hunink, Milton Weinstein, Eve Wittenberg, Michael Drummond, Joseph Pliskin, John Wong, and Paul Glasziou. 2014. *Decision Making in Health and Medicine: Integrating Evidence and Values.* Cambridge University Press. Retrieved from http://jrsm.rsmjournals.com/cgi/doi/10.1258/jrsm.95.2.108-a.

[88] IBM. 2015. Edgware fabric: A service bus for the physical world. Retrieved from https://goo.gl/CH4U6W.

[89] IBM. 2015. Smarter planet. Retrieved from https://goo.gl/vW0iLd.

[90] IEEE. 2013. 1888.3-2013—IEEE Standard for Ubiquitous Green Community Control Network: Security. Retrieved from http://ieeexplore.ieee.org/servlet/opac?punumber=6675753.

[91] International Telecommunication Union. 2012. *Overview of the Internet of Things.* Technical Report. International Telecommunication Union. Retrieved from http://www.itu.int/ITU-T/recommendations/rec.aspx?rec=11559.

[92] Internet Engineering Task Force. 1999. Simple service discovery protocol/1.0. Retrieved from https://tools.ietf.org/html/draft-cai-ssdp-v1-03.

[93] Internet Engineering Task Force. 2005. RFC 4301: Security architecture for the internet protocol. Retrieved from https://tools.ietf.org/html/rfc4301.

[94] Internet Engineering Task Force. 2013. RFC 6762: Multicast DNS. Retrieved from https://tools.ietf.org/html/rfc6762.

[95] Internet Engineering Task Force. 2014. The constrained application protocol (CoAP). Retrieved from https://tools.ietf.org/html/rfc7252.

[96] Internet Engineering Task Force. 2017. CoRE resource directory. Retrieved from https://tools.ietf.org/html/draft-ietf-core-resource-directory-11.

[97] Michael Isard, Mihai Budiu, Yuan Yu, Andrew Birrell, and Dennis Fetterly. 2007. Dryad: Distributed data-parallel programs from sequential building blocks. *ACM SIGOPS Operat. Syst. Rev.* 41, 3 (2007), 59–72. Retrieved from DOI: https://doi.org/10.1145/1272996.1273005

[98] Antonio Jara, Pablo Lopez, David Fernandez, Jose Castillo, Miguel Zamora, and Antonio Skarmeta. 2013. Mobile digcovery: A global service discovery for the internet of things. In *Proceedings of the 27th International Conference on Advanced Information Networking and Applications Workshops.* Retrieved from DOI: https://doi.org/10.1109/WAINA.2013.261

[99] Antonio J. Jara, Dominique Genoud, and Yann Bocchi. 2015. Big data for smart cities with KNIME a real experience in the smartsantander testbed. *Softw.: Practice Exper.* 45, 8 (Aug. 2015), 1145–1160. Retrieved from DOI: https://doi.org/10.1002/spe.2274 arxiv:1008.1900

[100] Jay Kreps. 2013. The Log: What every software engineer should know about real-time data's unifying abstraction. Retrieved from https://goo.gl/b07C4f.

[101] Simon Jirka and Daniel Nüst. 2010. *OGC Sensor Instance Registry Discussion Paper.* Technical Report. Open Geospatial Consortium. Retrieved from https://wiki.52north.org/SensorWeb/SensorInstanceRegistry.

[102] Robert Kallman, Hideaki Kimura, Jonathan Natkins, Andrew Pavlo, Alexander Rasin, Stanley Zdonik, Evan P. C. Jones, Samuel Madden, Michael Stonebraker, Yang Zhang, John Hugg, and Daniel J. Abadi. 2008. H-store: A high-performance, distributed main memory transaction processing system. *Proc. VLDB Endow.* 1, 2 (2008), 1496–1499. Retrieved from DOI: https://doi.org/10.1145/1454159.1454211

[103] Karthik Kambatla, Giorgos Kollias, Vipin Kumar, and Ananth Grama. 2014. Trends in big data analytics. *J. Parallel Distrib. Comput.* 74, 7 (2014), 2561–2573. Retrieved from DOI: https://doi.org/10.1016/j.jpdc.2014.01.003

[104] Andreas Kamilaris, Feng Gao, Francesc X. Prenafeta-Boldu, and Muhammad Intizar Ali. 2017. Agri-IoT: A semantic framework for internet of things-enabled smart farming applications. In *Proceedings of the 2016 IEEE 3rd World Forum on Internet of Things.* 442–447. Retrieved from DOI: https://doi.org/10.1109/WF-IoT.2016.7845467

[105] Lisa Kart. 2012. *Advancing Analytics.* Technical Report. Gartner Inc. Retrieved from http://meetings2.informs.org/analytics2013/AdvancingAnalytics.

[106] Daniel Keim, Gennady Andrienko, Jean-daniel Fekete, and Guy Melançon. 2008. Visual analytics: Definition, process, and challenges. In *Information Visualization.* Springer, 154–175. Retrieved from DOI: https://doi.org/10.1007/978-3-540-70956-5_7

[107] Daniel Keim, Jörn Kohlhammer, Geoffrey Ellis, and Florian Mansmann. 2010. *Mastering the Information Age Solving Problems with Visual Analytics.* EuroGraphics. Retrieved from DOI: https://doi.org/10.1016/j.procs.2011.12.035 arxiv:arXiv:1011.1669v3.

[108] Khalid S. Khan, Regina Kunz, Jos Kleijnen, and Gerd Antes. 2003. Five steps to conducting a systematic review. *J. Roy. Soc. Med.* 96, 3 (2003), 118–121. Retrieved from DOI: https://doi.org/10.1258/jrsm.96.3.118

[109] Shweta Khare, Kyoungho An, and Aniruddha Gokhale. 2015. Functional reactive stream processing for data-centric publish/subscribe systems. In *Proceedings of the 29th IEEE International Parallel & Distributed Processing Symposium.*

[110] Gerd Kortuem, Fahim Kawsar, Daniel Fitton, and Vasughi Sundramoorthy. 2010. Smart objects as building blocks for the internet of things. *IEEE Internet Comput.* 14 (2010), 44–51. Retrieved from DOI: https://doi.org/10.1109/MIC.2009.143

[111] Deanne Larson and Victor Chang. 2016. A review and future direction of agile, business intelligence, analytics and data science. *Int. J. Info. Manage.* 36, 5 (2016), 700–710. Retrieved from DOI: https://doi.org/10.1016/j.ijinfomgt.2016.04.013

[112] Steve Lavalle, Michael S. Hopkins, Eric Lesser, Rebecca Shockley, and Nina Kruschwitz. 2010. Analytics: The new path to value. *MIT Sloan Manage. Rev.* 51, 2 (2010), 1–24. Retrieved from https://www-935.ibm.com/services/uk/gbs/pdf/Analytics.

[113] Jung Hoon Lee, Marguerite Gong Hancock, and Mei Chih Hu. 2013. Towards an effective framework for building smart cities: Lessons from seoul and san francisco. *Technol. Forecast. Soc. Change* 89 (2013), 80–99. Retrieved from DOI: https://doi.org/10.1016/j.techfore.2013.08.033

[114] Shu Hsien Liao, Pei Hui Chu, and Pei Yuan Hsiao. 2012. Data mining techniques and applications—A decade review from 2000 to 2011. *Expert Syst. Appl.* 39, 12 (2012), 11303–11311. Retrieved from DOI: https://doi.org/10.1016/j.eswa.2012.02.063 arxiv:1202.1112

[115] Thomas Liebig, Nico Piatkowski, Christian Bockermann, and Katharina Morik. 2014. Predictive trip planning-smart routing in smart cities. In *Proceedings of the Workshops of the EDBT/ICDT 2014 Joint Conference.* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.429.2841.

[116] Jie Lin, Wei Yu, Nan Zhang, Xinyu Yang, Hanlin Zhang, and Wei Zhao. 2017. A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications. *IEEE Internet Things J.* 4, 5 (2017). Retrieved from DOI: https://doi.org/10.1109/JIOT.2017.2683200

[117] Honghai Liu, Shengyong Chen, and Naoyuki Kubota. 2013. Intelligent video systems and analytics: A survey. *IEEE Trans. Industr. Info.* 9, 3 (2013), 1222–1233. Retrieved from DOI: https://doi.org/10.1109/TII.2013.2255616

[118] Dave Locke. 2010. MQ Telemetry Transport (MQTT) V3.1 protocol specification. Retrieved from https://www.ibm.com/developerworks/library/ws-mqtt/index.html.

[119] Yucheng Low, Danny Bickson, Joseph Gonzalez, Carlos Guestrin, Aapo Kyrola, and Joseph M Hellerstein. 2012. Distributed GraphLab: A framework for machine learning and data mining in the cloud. *Proc. VLDB Endow.* 5, 8 (Apr. 2012), 716–727. Retrieved from DOI: https://doi.org/10.14778/2212351.2212354

[120] David Luckham. 2002. *The Power Of Events: An Introduction To Complex Event Processing In Distributed Enterprise Systems.* Addison-Wesley. Retrieved from DOI: https://doi.org/10.1007/978-3-540-88808-6_2

[121] Lustre. 2015. The Lustre Filesystem. Retrieved from http://lustre.opensfs.org/.

[122] Sam Madden. 2012. From databases to big data. *IEEE Internet Comput.* 16 (2012), 4–6. Retrieved from DOI: https://doi.org/10.1109/MIC.2012.50

[123] Ganapathy Mahalakshmi, Sridevi Sureshkumar, and S. Rajaram. 2016. A survey on forecasting of time-series data. In *Proceedings of the 2016 International Conference on Computing Technologies and Intelligent Data Engineering.* Retrieved from DOI: https://doi.org/10.1109/ICCTIDE.2016.7725358

[124] Chin Mak and Henry Fan. 2006. Heavy flow-based incident detection algorithm using information from two adjacent detector stations. *J. Intell. Transport. Syst.* 10, 1 (2006), 23–31. Retrieved from DOI: https://doi.org/10.1080/15472450500455229

[125] James Manyika, Michael Chui, Peter Bisson, Jonathan Woetzel, Richard Dobbs, Jacques Bughin, and Dan Aharon. 2015. *The Internet of Things: Mapping the Value Beyond the Hype.* Technical Report. McKinsey Global Institute. Retrieved from http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/the-internet-of-things-the-value-of-digitizing-the-physical-world.

[126] James Manyika, Michael Chui, and Jacques Bughin. 2013. *Disruptive Technologies: Advances That Will Transform Life, Business, And The Global Economy.* Technical Report. McKinsey Global Institute. Retrieved from https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/disruptive-technologies.

[127] MapR Technologies. 2014. *Stream Processing with MapR.* Technical Report. MapR Inc. Retrieved from https://mapr.com/resources/stream-processing-mapr/.

[128] Nathan Marz and James Warren. 2014. *Big Data: Principles and Best Practices of Scalable Real-time Data Systems.* Mannings. arxiv:1-933988-16-9.

[129] MemSQL Inc.2015. MemSQL. Retrieved from http://www.memsql.com/.
[130] Mesosphere. 2017. Mesosphere. Retrieved from https://mesosphere.com.
[131] Microsoft. 2015. Microsoft Azure. Retrieved from http://azure.microsoft.com/en-gb/.
[132] Milan Milenkovic. 2015. A case for interoperable iot sensor data and meta-data formats. *Ubiquity* 2015 (November 2015), 1–7. Retrieved from DOI: https://doi.org/10.1145/2822643
[133] Daniele Miorandi, Sabrina Sicari, Francesco De Pellegrini, and Imrich Chlamtac. 2012. Internet of things: Vision, applications and research challenges. *Ad Hoc Netw.* 10, 7 (2012), 1497–1516. Retrieved from DOI: https://doi.org/10.1016/j.adhoc.2012.02.016
[134] Sebastian Mittelstadt, Michael Behrisch, Stefan Weber, Tobias Schreck, Andreas Stoffel, Rene Pompl, Daniel Keim, Holger Last, and Leishi Zhang. 2012. Visual analytics for the big data era—A comparative review of state-of-the-art commercial systems. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology*. Retrieved from http://ieeexplore.ieee.org/document/6400554/.
[135] Arijit Mukherjee, Swarnava Dey, Himadri Sekhar Paul, and Batsayan Das. 2013. Utilising condor for data parallel analytics in an IoT context—An experience report. In *Proceedings of the 9th IEEE International Conference on Wireless and Mobile Computing, Networking and Communications*. Retrieved from DOI: https://doi.org/10.1109/WiMOB.2013.6673380
[136] Arijit Mukherjee, Arpan Pal, and Prateep Misra. 2012. Data analytics in ubiquitous sensor-based health information systems. In *Proceedings of the 6th International Conference on Next Generation Mobile Applications, Services, and Technologies*. Retrieved from DOI: https://doi.org/10.1109/NGMAST.2012.39
[137] Arijit Mukherjee, Himadri Sekhar Paul, Swarnava Dey, and Ansuman Banerjee. 2014. ANGELS for distributed analytics in IoT. In *Proceedings of IEEE World Forum on Internet of Things*. Retrieved from DOI: https://doi.org/10.1109/WF-IoT.2014.6803230
[138] Ujjal Kumar Mukherjee and Snigdhansu Chatterjee. 2014. Fast algorithm for computing weighted projection quantiles and data depth for high-dimensional large data clouds. In *Proceedings of the 2014 IEEE International Conference on Big Data*. Retrieved from http://ieeexplore.ieee.org/document/7004358/.
[139] Stefan Nastic, Sanjin Sehic, Michael Vögler, Hong Linh Truong, and Schahram Dustdar. 2013. PatRICIA—A novel programming model for IoT applications on cloud platforms. In *Proceedings of the 6th IEEE International Conference on Service-Oriented Computing and Applications*. Retrieved from DOI: https://doi.org/10.1109/SOCA.2013.48
[140] Septimiu Nechifor, Anca Petrescu, Dan Puiu, and Bogdan Tarnauca. 2014. Predictive analytics based on CEP for logistic of sensitive goods. In *Proceedings of the International Conference on Optimization of Electrical and Electronic Equipment*. Retrieved from http://ieeexplore.ieee.org/document/6850965/.
[141] David Niewolny. 2013. *How the Internet of Things Is Revolutionizing Healthcare.* Technical Report. Freescale Semiconductor. Retrieved from http://cache.freescale.com/files/corporate/doc/white.
[142] Office of National Statistics. 2013. *Population and Household Estimates for the United Kingdom.* Technical Report. Retrieved from https://goo.gl/dAUEjm.
[143] Niall O'Hara, Marco Slot, Dan Marinescu, Jan Čurn, Dawei Yang, Mikael Asplund, Mélanie Bouroche, Siobhán Clarke, and Vinny Cahill. 2012. MDDSVsim: An integrated traffic simulation platform for autonomous vehicle research. In *Proceedings of the International Workshop on Vehicular Traffic Management for Smart Cities.*
[144] Oxford English Dictionary. 2017. "analytics, n." Retrieved from http://www.oed.com/view/Entry/273413.
[145] Kasey Panetta. 2017. Top trends in the Gartner hype cycle for emerging technologies. Retrieved from http://www.gartner.com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017/.
[146] Panoply. 2017. Panoply smart data warehouse. Retrieved from https://panoply.io/.
[147] Paradigm4. 2014. *Leaving Data on the Table.* Technical Report. Retrieved from http://goo.gl/6vBhk3.
[148] Andrew Pavlo, Gustavo Angulo, Joy Arulraj, Haibin Lin, Jiexi Lin, Lin Ma, Prashanth Menon, Todd C Mowry, Matthew Perron, Ian Quah, Siddharth Santurkar, Anthony Tomasic, Skye Toor, Dana Van Aken, Ziqi Wang, Yingjun Wu, Ran Xian, and Tieying Zhang. 2017. Self-driving database management systems. In *Proceedings of the 8th Biennial Conference on Innovative Data Systems Research.* Retrieved from http://pelotondb.io/publications/.
[149] Fernando Pereira, John Lafferty, and Andrew Mccallum. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of 18th International Conference on Machine Learning*. Retrieved from http://dl.acm.org/citation.cfm?id=655813.
[150] Charith Perera, Arkady Zaslavsky, Peter Christen, Michael Compton, and Dimitrios Georgakopoulos. 2013. Context-aware sensor search, selection and ranking model for internet of things middleware. In *Proceedings of IEEE International Conference on Mobile Data Management.* Retrieved from DOI: https://doi.org/10.1109/MDM.2013.46
[151] Charith Perera, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos. 2014. Context aware computing for the internet of things: A survey. *IEEE Commun. Surveys Tutor.* 16, 1 (2014), 414–454. Retrieved from DOI: https://doi.org/10.1109/SURV.2013.042313.00197

[152] Christy Pettey. 2010. Gartner's 2010 hype cycle special report. Retrieved from http://www.gartner.com/newsroom/id/1447613.

[153] Christy Pettey and Laurence Goasduff. 2011. Gartner's 2011 hype cycle special report. Retrieved from http://www.gartner.com/newsroom/id/1763814.

[154] Christy Pettey and Rob van der Meulen. 2012. Gartner's 2012 hype cycle for emerging technologies. Retrieved from http://www.gartner.com/newsroom/id/2124315.

[155] Nicola Piovesan, Leo Turi, Enrico Toigo, Borja Martinez, and Michele Rossi. 2016. Data analytics for smart parking applications. *Sensors* 16, 10 (2016), 1–25. Retrieved from DOI: https://doi.org/10.3390/s16101575

[156] Pivotal Inc. 2015. Greenplum database. Retrieved from http://pivotal.io/big-data/pivotal-greenplum-database.

[157] Joern Ploennigs, Anika Schumann, and Freddy Lécué. 2014. Adapting semantic sensor networks for smart building diagnosis. In *Proceedings of the 13th International Semantic Web Conference*. Retrieved from DOI: https://doi.org/10.1007/978-3-319-11915-1_20

[158] Andrew Prunicki. 2009. *Apache Thrift*. Technical Report. Object Computing Inc. Retrieved from https://thrift.apache.org/.

[159] Bastian Quilitz and Ulf Leser. 2008. Querying distributed RDF data sources with SPARQL. In *Proceedings of the 5th European Semantic Web Conference*. Retrieved from DOI: https://doi.org/10.1007/978-3-540-68234-9_39

[160] Quobyte. 2017. Quobyte and XtreemFS. Retrieved from https://www.quobyte.com/containers.

[161] Ioan Raicu, Ian T. Foster, and Pete Beckman. 2011. Making a case for distributed file systems at exascale. In *Proceedings of the 3rd International Workshop on Large-scale System and Application Performance*. 11. Retrieved from DOI: https://doi.org/10.1145/1996029.1996034

[162] Ioan Raicu, Ian T. Foster, and Pete Beckman. 2012. Making a case for distributed file systems at exascale. In *Proceedings of the 3rd International Workshop On Large-scale System And Application Performance*. Retrieved from DOI: https://doi.org/10.1145/1996029.1996034

[163] Nur Aini Rakhmawati and Michael Hausenblas. 2012. On the impact of data distribution in federated SPARQL queries. In *Proceedings of 6th IEEE International Conference on Semantic Computing*. Retrieved from DOI: https://doi.org/10.1109/ICSC.2012.72

[164] Nur Aini Rakhmawati, Jürgen Umbrich, Marcel Karnstedt, Ali Hasnain, and Michael Hausenblas. 2013. *Querying Over Federated SPARQL Endpoints—A State of the Art Survey*. Technical Report. Digital Enterprise Research Institute. Retrieved from https://arxiv.org/abs/1306.1723.

[165] Partha Pratim Ray. 2015. Towards an internet of things based architectural framework for defence. In *Proceedings of the 2015 International Conference on Control Instrumentation Communication and Computational Technologies*. Retrieved from DOI: https://doi.org/10.1109/ICCICCT.2015.7475314

[166] Partha Pratim Ray. 2016. A survey on internet of things architectures. *J. King Saud Univ. Comput. Info. Sci.* 30, 3 (2016). Retrieved from DOI: https://doi.org/10.1016/j.jksuci.2016.10.003

[167] Ahmad Razip, Abish Malik, Shehzad Afzal, Matthew Potrawski, Ross Maciejewski, Yun Jang, Niklas Elmqvist, and David Ebert. 2014. A mobile visual analytics approach for law enforcement situation awareness. In *Proceedings of the 2014 IEEE Pacific Visualization Symposium*. Retrieved from DOI: https://doi.org/10.1109/PacificVis.2014.54

[168] Mohammad Abdur Razzaque, Marija Milojevic-Jevric, Andrei Palade, and Siobhán Cla. 2016. Middleware for internet of things: A survey. *IEEE Internet Things J.* 3, 1 (2016), 70–95. Retrieved from DOI: https://doi.org/10.1109/JIOT.2015.2498900

[169] Janessa Rivera and Rob Meulen. 2013. Gartner's 2013 hype cycle for emerging technologies. Retrieved from http://www.gartner.com/newsroom/id/2575515.

[170] Janessa Rivera and Rob Meulen. 2014. Gartner's 2014 hype cycle for emerging technologies. Retrieved from http://www.gartner.com/newsroom/id/2819918.

[171] Janessa Rivera and Rob Meulen. 2015. Gartner's 2015 hype cycle for emerging technologies. Retrieved from http://www.gartner.com/newsroom/id/3114217.

[172] Silva Robak, Bogdan Franczyk, and Marcin Robak. 2013. Applying big data and linked data concepts in supply chains management. In *Proceedings of the Federated Conference on Computer Science and Information Systems*. Retrieved from http://ieeexplore.ieee.org/document/6644169/.

[173] Seref Sagiroglu and Duygu Sinanc. 2013. Big data: A review. In *International Conference on Collaboration Technologies and Systems*. Retrieved from DOI: https://doi.org/10.1109/CTS.2013.6567202

[174] Muhammad Saleem, Yasar Khan, Ali Hasnain, Ivan Ermilov, and Axel-Cyrille Ngonga Ngomo. 2014. A fine-grained evaluation of SPARQL endpoint federation systems. *Semant. Web J.* 1 (2014), 1–5. Retrieved from http://www.semantic-web-journal.net/system/files/swj625.pdf.

[175] Rosario Salpietro, Luca Bedogni, Marco Di Felice, and Luciano Bononi. 2015. Park here! A smart parking system based on smartphones' embedded sensors and short range communication technologies. In *Proceedings of the 2015 IEEE World Forum on Internet of Things*. Retrieved from DOI: https://doi.org/10.1109/WF-IoT.2015.7389020

[176] Luis Sanchez, Jose Antonio Galache, Veronica Gutierrez, Jose Manuel Hernandez, Jesus Bernat, Alex Gluhak, and Tomas Garcia. 2011. SmartSantander: The meeting point between future internet research and experimentation and the smart cities. In *Proceedings of the Future Network & Mobile Summit.* Retrieved from http://ieeexplore.ieee.org/document/6095264/.

[177] Mehadev Satyanarayanan, Paramvir Bahl, Ramon Caceres, and Nigel Davies. 2009. The case for VM-base cloudlets in mobile computing. *Pervas. Comput.* 8 (2009), 14–23. Retrieved from DOI: https://doi.org/10.1109/MPRV.2009.82

[178] Francois Schnizler, Thomas Liebig, Shie Mannor, Gustavo Souto, Sebastian Bothe, and Hendrik Stange. 2014. Heterogeneous stream processing for disaster detection and alarming. In *Proceedings of the 2014 IEEE International Conference on Big Data*. Retrieved from http://ieeexplore.ieee.org/document/7004323/.

[179] Jurgen Schonwalder, Martin Bjorklund, and Phil Shafer. 2010. Network configuration management using NETCONF and YANG. *IEEE Commun. Mag.* 48, 9 (2010), 166–173. Retrieved from DOI: https://doi.org/10.1109/MCOM.2010.5560601

[180] Andreas Schwarte, Peter Haase, Katja Hose, Ralf Schenkel, and Michael Schmidt. 2011. FedX: Optimization techniques for federated query processing on linked data. In *Proceedings of the 10th International Semantic Web Conference.* Retrieved from DOI: https://doi.org/10.1007/978-3-642-25073-6_38

[181] Pallavi Sethi and Smruti R. Sarangi. 2017. Internet of things: Architectures, protocols, and applications. *J. Electric. Comput. Eng.* 2017 (2017). Retrieved from DOI: https://doi.org/10.1155/2017/9324035

[182] Rajeev Sharma, Peter Reynolds, Rens Scheepers, Peter B. Seddon, and Graeme G. Shanks. 2010. Business analytics and competitive advantage: A review and a research agenda. In *Bridging the Socio-technical Gap in Decision Support Systems: Challenges for the Next Decade.* IOS Press, 187–198.

[183] Galit Shmueli, Peter C. Bruce, Inbal Yahav, Nitin R. Patel, and Kenneth C. Lichtendahl Jr. 2017. *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R.* John Wiley & Sons. Retrieved from DOI: https://doi.org/978-1-118-87936-8

[184] Galit Shmueli and Otto Koppiu. 2011. Predictive analytics in information systems research. *MIS Quarterly* 35, 3 (2011), 553–572. Retrieved from https://doi.org/10.2307/23042796.

[185] Roman Y. Shtykh and Toshihiro Suzuki. 2014. Distributed data stream processing with Onix. In *Proceedings of the 4th IEEE International Conference on Big Data and Cloud Computing.* Retrieved from DOI: https://doi.org/10.1109/BDCloud.2014.54

[186] Steve Sill, Blake Christie, Ann Diephaus, Dan Garretson, Kay Sullivan, and Susan Sloan. 2011. *Intelligent Transportation Systems (ITS) Standards Program Strategic Plan.* Technical Report. U.S. Department of Transportation.

[187] Eugene Siow, Thanassis Tiropanis, and Wendy Hall. 2017. Ewya: An interoperable fog computing infrastructure with RDF stream processing. In *Proceedings of the 4th International Conference on Internet Science.* Retrieved from https://eprints.soton.ac.uk/412749/.

[188] John A. Stankovic. 2014. Research directions for the internet of things. *IEEE Internet Things J.* 1, 1 (2014), 3–9. Retrieved from DOI: https://doi.org/10.1109/JIOT.2014.2312291

[189] Guo-Dao Sun, Ying-Cai Wu, Rong-Hua Liang, and Shi-Xia Liu. 2013. A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *J. Comput. Sci. Technol.* 28, 5 (2013), 852–867. Retrieved from DOI: https://doi.org/10.1007/s11390-013-1383-8

[190] Teradata. 2015. Teradata database. Retrieved from http://goo.gl/hLPwIV.

[191] Andrea Tosatto, Pietro Ruiu, and Antonio Attanasio. 2015. Container-based orchestration in cloud: State of the art and challenges. In *Proceedings of the 9th International Conference on Complex, Intelligent and Software Intensive Systems.* IEEE. Retrieved from DOI: https://doi.org/10.1109/CISIS.2015.35

[192] John W. Tukey. 1962. The future of data analysis. *Ann. Math. Stat.* 33, 1 (1962), 1–67. Retrieved from DOI: https://doi.org/10.1214/aoms/1177704711

[193] Efraim Turban, Ramesh Sharda, and Dursun Delen. 2014. *Businesss Intelligence and Analytics: Systems for Decision Support.* Pearson. Retrieved from http://catalogue.pearsoned.co.uk/educator/product/Business-Intelligence-and-Analytics-Systems-for-Decision-Support-Global-Edition/9781292009209.page.

[194] Ubeam. 2017. ubeam. Retrieved from http://ubeam.com/.

[195] Ellen van Nunen, Maurice Kwakkernaat, Jeroen Ploeg, and Bart Netten. 2012. Cooperative competition for future mobility. *IEEE Trans. Intell. Transport. Syst.* 13, 3 (2012), 1018–1025. Retrieved from DOI: https://doi.org/10.1109/TITS.2012.2200475

[196] Rajesh Vargheese and Hazim Dahir. 2014. An IoT/IoE enabled architecture framework for precision on shelf availability. In *Proceedings of the IEEE International Conference on Big Data.* Retrieved from http://ieeexplore.ieee.org/document/7004418.

[197] Hal Varian. 2009. How the web challenges managers. Retrieved from http://www.mckinsey.com/industries/high-tech/our-insights/hal-varian-on-how-the-web-challenges-managers.

[198] Cor Verdouw, Adrie Beulens, and Jack van der Vorst. 2013. Virtualisation of floricultural supply chains: A review from an IoT perspective. *Comput. Electron. Agric.* 99 (2013), 160–175. Retrieved from DOI: https://doi.org/10.1016/j.compag.2013.09.006

[199] Ovidiu Vermesan and Peter Friess. 2013. *Internet of Things: Converging Technologies for Smart Environments and Integrated Ecosystems.* River Publishers. Retrieved from DOI: https://doi.org/10.2139/ssrn.2324902

[200] Ovidiu Vermesan and Peter Friess. 2014. *Internet of Things: From Research and Innovation to Market Deployment.* Vol. 6. River Publishers. Retrieved from https://www.riverpublishers.com/book.

[201] Massimo Villari, Antonio Celesti, and Maria Fazio. 2014. AllJoyn Lambda: An architecture for the management of smart environments in IoT. In *Proceedings of 2014 International Conference on Smart Computing Workshops.* Retrieved from http://ieeexplore.ieee.org/document/7046676/.

[202] Mark Walport. 2014. *The Internet of Things: Making the Most of the Second Digital Revolution.* Technical Report. The United Kingdom Government Office for Science. Retrieved from https://www.gov.uk/government/publications/internet-of-things-blackett-review.

[203] Xin Wang, Thanassis Tiropanis, and Hugh C. Davis. 2013. LHD: Optimising linked data query processing using parallelisation. In *Proceedings of the Workshop on Linked Data on the Web.* Retrieved from https://eprints.soton.ac.uk/350719/.

[204] Sage A. Weil, Scott A. Brandt, Ethan L. Miller, and Darrell D. E. Long. 2006. Ceph: A scalable, high-performance distributed file system. In *Proceedings of the 7th Symposium on Operating Systems Design and Implementation.* Retrieved from https://dl.acm.org/citation.cfm?id=1298485.

[205] Marilyn Wolf. 2017. The physics of event-driven IoT systems. *IEEE Design Test* 34, 2 (2017), 87–90. Retrieved from DOI: https://doi.org/10.1109/MDAT.2016.2631082

[206] World Economic Forum. 2012. *The Global Information Technology Report 2012 Living in a Hyperconnected World.* Technical Report.

[207] Reynold S. Xin, Joseph E. Gonzalez, Michael J. Franklin, and Ion Stoica. 2013. GraphX: A resilient distributed graph system on spark. In *Proceedings of the 1st International Workshop on Graph Data Management Experiences and Systems.* ACM Press, New York, New York. Retrieved from DOI: https://doi.org/10.1145/2484425.2484427 arxiv: 1402.2394.

[208] Lida Xu, Wu He, and Shancang Li. 2014. Internet of things in industries: A survey. *IEEE Trans. Industr. Informat.* 4 (2014), 1–11. Retrieved from DOI: https://doi.org/10.1109/TII.2014.2300753

[209] Xiaomin Xu, Sheng Huang, Yaoliang Chen, Kevin Brown, Inge Halilovic, and Wei Lu. 2014. TSaaaS: Time-series analytics as a service on IoT. In *Proceedings of the IEEE International Conference on Web Services.* Retrieved from DOI: https://doi.org/10.1109/ICWS.2014.45

[210] Zheng Xu, Yunhuai Liu, Hui Zhang, Xiangfeng Luo, Lin Mei, and Chuanping Hu. 2017. Building the multi-modal storytelling of urban emergency events based on crowdsensing of social media analytics. *Mobile Netw. Appl.* 22, 2 (2017), 218–227. Retrieved from DOI: https://doi.org/10.1007/s11036-016-0789-2

[211] Fan Yang, Nelson Matthys, Rafael Bachiller, Sam Michiels, Wouter Joosen, and Danny Hughes. 2015. uPnP: Plug and play peripherals for the internet of things. In *Proceedings of the 10th European Conference on Computer Systems.* Retrieved from DOI: https://doi.org/10.1145/2741948.2741980

[212] Feng Ye, Zhi-Jian Wang, Fa-Chao Zhou, Ya-Pu Wang, and Yuan-Chao Zhou. 2013. Cloud-based big data mining & analyzing services platform integrating R. In *Proceedings of the 2013 International Conference on Advanced Cloud and Big Data.* Retrieved from DOI: https://doi.org/10.1109/CBD.2013.13

[213] Jennifer Yick, Biswanath Mukherjee, and Dipak Ghosal. 2008. Wireless sensor network survey. *Comput. Netw.* 52 (2008), 2292–2330. Retrieved from DOI: https://doi.org/10.1016/j.comnet.2008.04.002

[214] Jian Yin, Anand Kulkarni, Sumit Purohit, Ian Gorton, and Bora Akyol. 2011. Scalable real-time data management for smart grid. In *Proceedings of the Middleware 2011 Industry Track Workshop.* Retrieved from DOI: https://doi.org/10.1145/2090181.2090182

[215] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, and Ankur Dave. 2012. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation.* Retrieved from DOI: https://doi.org/10.1111/j.1095-8649.2005.00662.x

[216] Andrea Zanella, Nicola Bui, Angelo P. Castellani, Lorenzo Vangelista, and Michele Zorzi. 2014. Internet of things for smart cities. *IEEE Internet Things J.* 1, 1 (2014), 22–32. Retrieved from DOI: https://doi.org/10.1109/JIOT.2014.2306328

[217] Zhihua Zhou, Nitesh V. Chawla, Yaochu Jin, and Graham J. Williams. 2014. Big data opportunities and challenges: Discussions from data analytics perspectives. *IEEE Comput. Intell. Mag.* 9, 4 (2014), 62–74. Retrieved from DOI: https://doi.org/10.1109/MCI.2014.2350953

[218] Holger Ziekow and Zbigniew Jerzak. 2014. The DEBS 2014 grand challenge. *Proceedings of the 8th ACM International Conference on Distributed Event-based Systems*. Retrieved from DOI: https://doi.org/10.1145/2611286.2611333

[219] Leishi Zhang, Andreas Stoffel, Michael Behrisch, Sebastian Mittelstadt, Tobias Schreck, René Pompl, Stefan Weber, Holger Last, and Daniel Keim. 2012. Visual analytics for the big data era—A comparative review of state-of-the-art commercial systems. In *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology*. Retrieved from DOI: https://doi.org/10.1109/VAST.2012.6400554

Received May 2016; revised October 2017; accepted April 2018