PART I

IoT Analytics Enablers

1

Introducing IoT Analytics

John Soldatos

Athens Information Technology, Greece

1.1 Introduction

The internet-of-things (IoT) paradigm represents one of the next evolutionary steps in internet-based computing, which is already having a positive impact in a large number of application domains including smart cities, sustainable living, healthcare, manufacturing and more. IoT analytics refers to the analysis of data from multiple IoT data sources, including sensors, actuators, smart devices and other internet connected objects. The collection and analysis of data streams from IoT sources is nowadays considered a key element of the IoT's disruptive power, as well as a prerequisite to realizing IoT's hyped market potential. Indeed, according to a recent report by McKinsey [1], less than 1% of IoT data is currently used, which is a serious set-back to maximizing IoT's business value. For example, most IoT analytics applications are nowadays used for anomaly detection and control rather than for optimization and prediction, which are the applications that will provide the greatest business value in the coming years.

1.2 IoT Data and BigData

The rise of future internet technologies, including cloud computing and BigData analytics, enables the wider deployment and use of sophisticated IoT analytics applications, beyond simple sensor processing applications. It is therefore no accident that IoT technologies are converging with cloud computing and BigData analytics technologies towards creating and deploying advanced applications that process IoT streams.

The integration of IoT data streams within cloud computing infrastructures enables IoT analytics applications to benefit from the capacity, performance

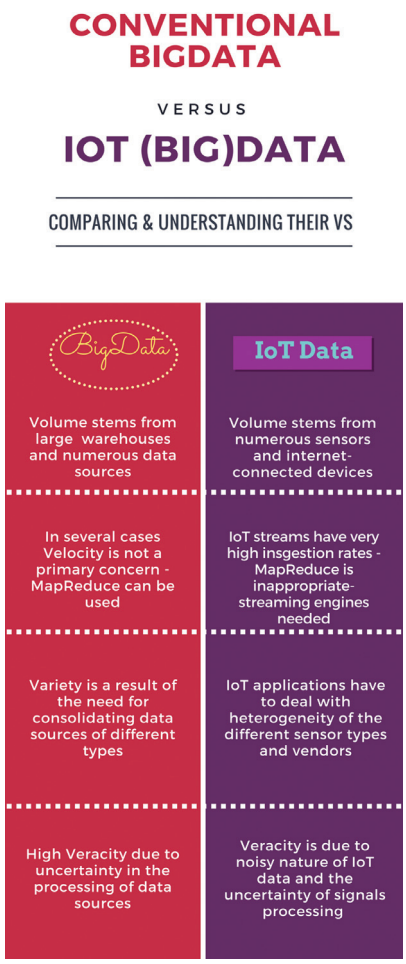

Figure 1.1 The Vs of BigData and IoT (Big)Data.

and scalability of cloud computing infrastructures. In several cases, IoT analytics applications are also integrated with edge computing infrastructures, which decentralize processing of IoT data streams at the very edge of the network, while transferring only selected IoT data from the edge devices to the cloud. Therefore, it is very common to deploy IoT analytics applications within edge and/or cloud computing infrastructures.

In addition to the affiliation between IoT analytics and cloud computing infrastructures, there is a close relation between IoT analytics with BigData analytics. Indeed, IoT data are essentially BigData since they feature several of the Vs of BigData, including (Figure 1.1):

- **Volume**: IoT data sources (such as sensors) produce in most cases very large volumes of data, which typically exceed the storage and processing capabilities of conventional database systems.
- **Velocity**: IoT data streams have commonly very high ingestion rates, as they are produced continually, in very high frequencies and in several times in very short timescales.
- **Variety**: Due to the large diversity of IoT devices, IoT data sources can be very heterogeneous both in terms of semantics and data formats.
- **Veracity**: IoT data are a classical example of noise data, which are characterized by uncertainty.

Therefore, systems, tools and techniques for developing and deploying BigData applications (including databases, data warehouses, streaming middleware and engines, data mining techniques and BigData developments tools), provide a good starting point for dealing with IoT analytics. However, IoT data and IoT analytics applications have in most cases to deal with their own peculiar challenges, which are not always common to the challenges of high volume, high velocity transactional applications. The tools and techniques that are discussed in this book are focused on the challenges of IoT data and IoT analytics applications, which are outlined in the following paragraph.

1.3 Challenges of IoT Analytics Applications

The main challenges associated with the development and deployment of IoT analytics applications are (Figure 1.2):

- **The heterogeneity of IoT data streams**: IoT data streams tend to be multi-modal and heterogeneous in terms of their formats, semantics and velocities. Hence, IoT analytics applications expose typically variety and veracity. BigData technologies provide the means for dealing with this heterogeneity in the scope of operationalized applications. However, accessing IoT data sources (including sensors and other types of internet connected devices) requires drivers and connectors, beyond what is typically deployed in transactional BigData applications (e.g., database drivers). Furthermore, dealing with semantic interoperability of diverse data streams requires techniques beyond the (syntactic) homogenization of data formats.
- **The varying data quality**: Several IoT streams are noisy and incomplete, which creates uncertainty in the scope of IoT analytics applications.

Statistical and probabilistic approaches must be therefore employed in order to take into account the noisy nature of IoT data streams, especially in cases where they stem from unreliable sensors. Also, different IoT data streams can be typically associated with different reliability, which should be considered in the scope of their integration in IoT analytics applications.

- **The real-time nature of IoT datasets**: IoT streams feature high velocities and for several application must be processed nearly in real-time. Hence, IoT analytics can greatly benefit from data streaming platforms, which are part of the BigData ecosystem. IoT devices (e.g., sensors) provide typically high-velocity data, which however can be in several cases controlled by focusing only on changes in data patterns and reports, rather than dealing with all the observations that stem from a given sensor.
- **The time and location dependencies of IoT streams**: IoT data come with temporal and spatial information, which is directly associated with their business value in a given application context. Hence, IoT analytics applications must in several cases process data in a timely fashion and from proper locations. Cloud computing techniques (including edge computing architectures) can greatly facilitate timely processing of information from given locations in the scope of large scale deployments. Note also that the spatial and temporal dimensions of IoT data can serve as a basis for dynamically selecting and filtering streams towards analytics applications for certain timelines and locations.
- **Privacy and security sensitivity**: IoT data are typically associated with stringent security requirements and privacy sensitivities, especially in the case of IoT applications that involve the collection and processing of personal data. Hence, IoT analytics need to be supported by privacy preservation techniques, such as the anonymization of personal data, as well as techniques for encrypted and secure data storage.
- **Data bias**: As in the majority of data mining problems, IoT datasets can lead to biased processing and hence a thorough understanding and scrutiny of both training and test datasets is required prior to their operationalized deployment. To this end, classical data mining techniques can also be applied in the IoT case. Note that the specification and deployment of IoT analytics systems entails techniques similar to those deployed in classical data mining problems, including the understanding of the data, the preparation of the data, the testing of data mining techniques and ultimately the development and deployment of a system that yields the desired performance and efficiency.

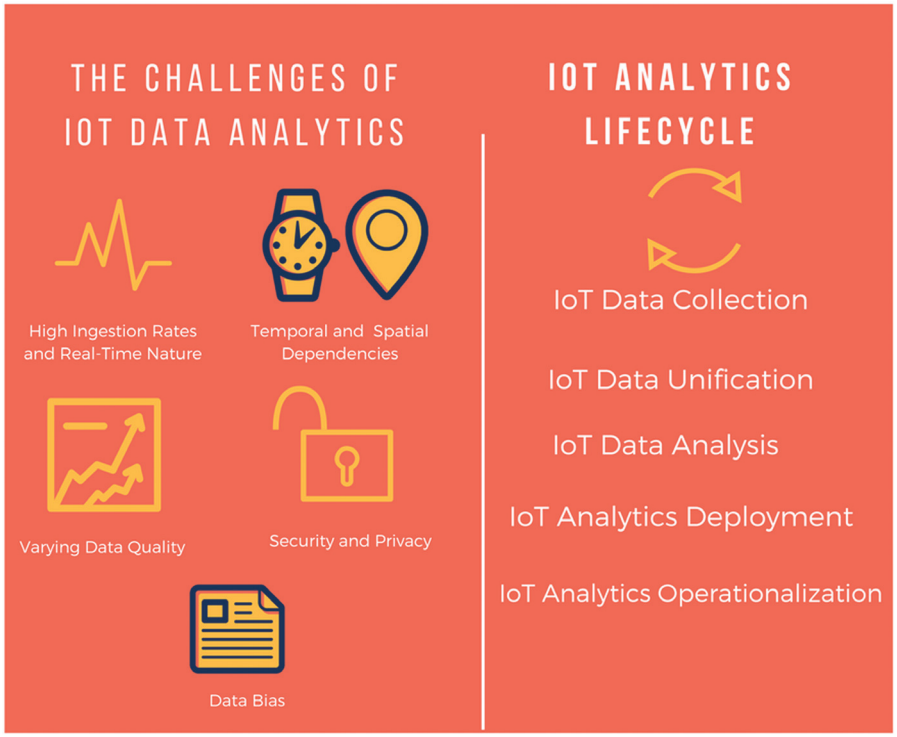

Figure 1.2 Main challenges and lifecycle phases of IoT analytics.

These challenges are evident in the IoT analytics lifecycle, which comprises a series of steps from data acquisition to analysis and visualization. This lifecycle is supported by cloud computing and BigData technologies, including data mining, statistical computing and scalable databases technology.

1.4 IoT Analytics Lifecycle and Techniques

The IoT analytics lifecycle comprises the phases of data collection, analysis and reuse. In particular:

- **1st Phase – IoT Data Collection**: As part of this phase IoT data are collected and enriched with the proper contextual metadata, such as location information and timestamps. Moreover, the data are validated in terms of their format and source of origin. Also, they are validated in terms of their integrity, accuracy and consistency. Hence, this phase

addresses several IoT analytics challenges, such as the need to ensure consistency and quality. Note that IoT data collection presents several peculiarities, when compared to traditional data consolidation of distributed data sources, such as the need to deal with heterogeneous IoT streams.

- **2nd Phase – IoT Data Analysis**: This phase deals with the structuring, storage and ultimate analysis of IoT data streams. The latter analysis involves the employment of data mining and machine learning techniques such as classification, clustering and rules mining. These techniques are typically used to transform IoT data to actionable knowledge.
- **3rd Phase – IoT Data Deployment, Operationalization and Reuse**: As part of this phase, the IoT analytics techniques identified in the previous steps are actually deployed, thus becoming operational. This phase ensures also the visualization of the IoT data/knowledge according to the needs of the application. Moreover, it enables the reuse of IoT knowledge and datasets across different applications.

These lifecycle phases are used in order to organize the development and deployment of IoT analytics systems. They can also serve as a basis for characterizing the maturity of IoT analytics deployments. As a prominent example, they can be used to analyze the level of "smartness" of a city, given that the maturity of a city is directly related to the sophistication of its analytics, but also to its ability to repurpose and reuse datasets and data analytics functions.

The tasks outlined in the above listed phases are supported by a range of data management and analysis disciplines, including:

- **IoT middleware and interoperability technologies**, which provide the means for collecting, structuring and unifying IoT data streams, thus addressing the variety and veracity challenges of IoT data.
- **Statistics**, which provide the theory for testing hypotheses about various insights stemming from IoT data.
- **Machine learning**, which enables the implementation of learning agents based on IoT data mining. Machine learning includes several heuristic techniques. The practical cases studies at the second part of the book make use of various machine learning schemes.
- **Data mining and Knowledge Discovery**, which combines theory and heuristics towards extracting knowledge. To this end, data cleaning, learning and visualization might be also employed.
- **Database management systems**, including Relational Database Management Systems (RDMS), NoSQL databases, BigData databases (such

as the HDFS (Hadoop Distributed File System), which provide the means for data persistence and management. Most of the practical examples and case studies presented in the book make use of some sort of database management systems in order to persist and manage the data.

- **Data streams management systems**, which handle transient streams, including continuous queries, while being able to handle data with very high ingestion rates, including streams featuring unpredictable arrival times and characteristics. IoT streaming systems are also supported by scalable, distributed data management systems.

The techniques that are outlined as part of subsequent chapters of this book use and in several cases enhance the above-listed techniques and systems for data collection, management and analysis. For example, the following chapters make direct references to distributed real-time streaming and event processing systems like Apache Spark[1] and Apache Storm[2]. Apache Storm is a free and open source distributed real-time computation system. It facilitates reliable processing of unbounded streams of data and deals with Real-time processing much in the same way Apache Hadoop deals with batch processing. Similarly, Apache Storm is an open source software that defines a broader set of operations when comparing to Hadoop, including transformation and actions which can be arbitrarily combined in any order. Spark supports several programming Languages including Java, Scala and Python. Note that the choice between Spark or Storm for IoT streaming and analytics can be based on a number of different factors. Spark is usually a good choice for projects using existing Hadoop or Mesos clusters, as well as for projects involving considerable graph processing, SQL access, or batch processing. Moreover, Spark provides a shell for interactive processing (something missing from Storm). On the other hand, Storm is a good choice for projects primarily focused on stream processing and Complex Event Processing that have structures matching Storm's capabilities. Storm provides boader language support, including support for the R language which is extremely popular among data scientists. Beyond Apache Spark and Storm projects, Apache Flink[3] is another open source stream processing framework, which can support low latency, high throughput, stateful and distributed processing for IoT data. It provides low-latency streaming ETL (Extract-Transform-Load) operations, offering much higher performance than traditional ETL for batch datasets. Moreover,

[1] spark.apache.org

[2] storm.apache.org

[3] flink.apache.org

Flink is event-time aware: Events stemming from the same real-world activity could arrive out of order in a Flink streaming system, but even in such cases Flink can maintain the order. Flink is a more recent project comparing to Spark, but it constantly gaining momentum in the industry due to its innovative and high-performance functionalities. Likewise the applications that are illustrated in the second part of the book take also advantage of these techniques. For example, several of the presented applications exploit NoSQL databases (such as MongoDB[4] and CouchDB[5] for data storage and management, while most of the applications deploy also some data mining method like classification, prediction and mining of association rules.

1.5 Conclusions

This introductory chapter has defined the scope of IoT analytics and presented related technologies. It has also outlined the close affiliation of IoT analytics with the cloud computing and BigData techniques. Furthermore, it has presented the main challenges of IoT analytics applications, which stem primarily from the unique characteristics and nature of IoT data. The rest of the book is destined to present technology solutions to these challenges, along with practical applications and case studies, which make use of such solutions. The presented solutions build in several cases over state-of-the-art IoT, cloud computing and BigData solutions, given that the integration of these technologies tends to become a norm for the variety of IoT analytics applications. The integration of IoT, cloud computing and BigData infrastructures and technologies is therefore the topic discussed in the next chapter.

[4] https://www.mongodb.com/

[5] http://couchdb.apache.org/

References

[1] Manyika, J., Chui, M., Bisson, P., Woetzel, J., Dobbs, R. Bughin, J., Aharon, D. *Unlocking the Potential of the Internet of Things*. McKinsey Global Institute, June (2015).

[2] J. Soldatos, et al. "*IoT analytics: Collect, Process, Analyze, and Present Massive Amounts of Operational data – Research and Innovation Challenges*" Chapter 7 in Book, "Building the Hyperconnected Society – IoT Research and Innovation Value Chains, Ecosystems and Markets", IERC Cluster Book 2015, River Publishers.