

Low-cost IoT, Big Data, and Cloud Platform for Developing Countries

Corentin Dupont¹, Mehdi Sheikhalishahi², Abdur Rahim¹, Congduc Pham³

¹ FBK/Create-Net

cduPont@fbk.eu, arahim@fbk.eu

² Innotec 21

mehdi.sheikhalishahi@innotec21.de

³ University of Pau

congduc.pham@univ-pau.fr

Abstract. Gartner forecasts that 6.4 billion connected things will be in use worldwide in 2016, up 30 percent from 2015, and will reach 20.8 billion by 2020. In 2016, 5.5 million new things will get connected every day. Furthermore, the current research and market trends shows the convergence between IoT and Big Data. On the other hand, Africa as a continent has seen very little of this activity. In this paper we present WAZIUP, a project aiming at building an open innovation platform able to accelerate innovation in rural Africa. The WAZIUP platform will allow to develop IoT applications coupled with Big Data capacities. The platform is tailored to the specific requirements and constraints of African users. We will give an overview of the WAZIUP IoT and Big Data platform, detail its technical aspects and finally introduce four use case deployments.

1 Introduction

TODO ► change this intro to fit paper objectives: low cost HW review, Security/AuthN/AuthZ, data analytics using Elastic search, market opportunities ◀

TODO ► C. Pham: the idea is to say that there will be a huge uptake of IoT devices that will be deployed worldwide for a large variety of applications; and that needs suitable platforms. This uptake will definitely come from the possibility and the availability of low-cost IoT devices that can be deployed in a very simple manner thanks to new radio technologies. ◀

ICT developments in Africa has already enabled significant modernizations across traditional sectors. Notable examples are the micro-health insurance accessible through mobile devices, index-based crop insurance and crowd-sourced management of public services. These innovative applications recognize and leverage commonalities between sectors, blur traditional lines, and open up a new field of opportunities.

The opportunity for ICT in Africa is huge especially for IoT and big data: those technologies are promising a big wave of innovation for our daily lives. The promise of IoT is to connect billions of sensors, devices, equipment and systems. In turn, the challenge is to drive business outcomes, consumer benefits, and to

create new value. The new mantras for the IoT Era is the collection, convergence and exploitation of data. The information is collected from sensors, devices, gateways, edge equipment and networks and stored in their respective IoT platforms. This information is processed in order to increase business efficiency through automation while reducing downtime and improving people productivity.

While developed countries are discussing about massive deployment of IoT, countries in Sub-Saharan Africa are still far from being ready to enjoy the full benefit of IoT. They face many challenges, such as the lack of infrastructure and the high cost platforms complexity in deployment. At the same time, it is urgent to promote IoT worldwide : WAZIUP will contribute by reducing part of the technology gap between EU and Africa. The goal of WAZIUP is to deploy an IoT and big data platform for African needs and validate it through several Sub Saharan Africa real-life use cases.

WAZIUP targets the rural community in Sub-Saharan Africa: about 64% of the population is living outside cities. The region will be predominantly rural for at least another generation. The pace of urbanization here is slower compared to other continents, and the rural population is expected to grow until 2045. The majority of rural residents live on less than few euros per day. Rural development is particularly imperative in sub-Saharan Africa, where half of the rural people depends on the agriculture/micro and small farm business, other half faces rare formal employment and pervasive unemployment. For rural development, technologies have to support several key application sectors such as living quality, health, agriculture and climate changes.

The biggest challenge of WAZIUP is to reduce costs and power consumption while increasing the robustness of the hardware. Hardware has to be robust enough so as to require lower maintenance and handle environmental and deployment treats as well. WAZIUP will present an innovative design of the IoT platform dedicated to the rural ecosystem. To achieve that, low-cost, generic building blocks will be deployed for maximum adaptation to end-applications in the context of the rural economy in developing countries. Another challenge of WAZIUP is to be able to manage the network deployment and to facilitate IoT communication. Lower cost solutions has to be used : privilege price and single hop dedicated communication networks, energy autonomous, with low maintenance costs and long lasting operations. Dynamic management of long range connectivity has to be taken into account (e.g., cope with network & service fluctuations), such as devices identification, abstraction/virtualization of devices, communication and network resources optimization. Finally, WAZIUP aim to exploit the potential of big-data applications in the specific rural context. Data will be collected from the IoT sensors themselves, but WAZIUP will also collect open data from other sources to build predictive models and enrich the platform.

From a technical standpoint, WAZIUP will pay attention to all related privacy and security aspects with specifics addressing the involved communities (farmers, developers).

Continued Openness will be ensured through the release of open specification and open software components and/or algorithms. Low-cost and low-energy consumption will be possible through the design of innovation hardware (sensors/actuators), and of IoT communication & network infrastructure.

The challenges outlined above will be tackled using an open IoT-Big Data Platform with affordable sensors connected through an IoT-Cloud open platform. The technical functionalities encompassed by the platform will be a cloud-based real-time data collection combined with analytics and automation software, an intelligent analytics of sensor and device data, an integration to third parties platforms and a Platform-as-a-Service (PaaS) provider. The PaaS will provide to business clientele with independently maintained platform upon which their web application, services and mobile applications can be built.

The rest of this book chapter is structured as follows: Section ?? presents the architecture of the WAZIUP platform. Section ?? shows the implementation chosen. Section ?? details the deployment of WAZIUP, constrained by its environment. Section ?? presents four use cases that will be used to validate the WAZIUP concepts. Section 6 presents a survey of the literature on the topic, together with a survey of the Big Data Open Source tools. Finally we conclude the chapter in Section 7.

2 The IoT era with low-cost IoT for all

2.1 IoT connectivity made easy

Recent Low-Power Wide Area Networks (LPWAN) technologies for Internet-of-Things (IoT) introduced by Sigfox and Semtech's (LoRaTM) are currently gaining incredible interest and are under intense deployment campaigns worldwide. They definitely initiated a new innovation cycle as they obviously provide a much better connectivity answer for IoT (most of IoT devices have small amount of data to send and very limited battery power) compared to traditional cellular-based connectivity (e.g. GSM/GPRS/3G) or short-range technologies such as IEEE 802.15.4. They offer several kilometers range without relay nodes to reach a central gateway, thus greatly simplifying large-scale deployment of IoT devices as opposed to the complex multi-hop approach needed by short-range radio technologies. Fig. 1 shows a typical extreme long-range 1-hop connectivity scenario to a long-range gateway, which is the single interface to Internet servers, using low-cost LoRa radio modules available from many vendors. Most of these long-range technologies can achieve 20km or higher range in line-of-sight (LOS) condition and about 2km-4km in non-LOS conditions [?, ?] such as in dense urban/city environments.

2.2 Low-cost DIY IoT hardware

Commercial IoT devices are getting mature but they are definitely too expensive for very low-income countries. In addition, these highly integrated devices are

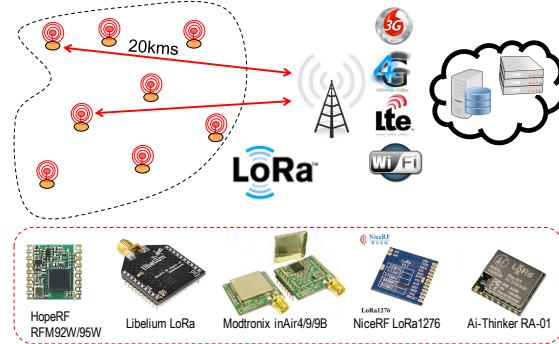


Fig. 1. Extreme long-range application with new radio technologies

difficult to repair with their parts being hardly locally replaced. The availability of low-cost, open-source hardware platforms such as Arduino boards definitely pushes for a Do-It-Yourself (DIY) and "off-the-shelves" design approach for a large variety of IoT applications. The Arduino ecosystem is large and proposes various board models, from large and powerful prototyping boards to smaller and less energy-consuming boards for final integration purposes as illustrated in Figure 2. For instance, the small form factor Arduino Pro Mini board based on an ATmega328 microcontroller has a high performance/price tradeoff and can be used to build a low-cost generic sensing IoT platform with LoRa long-range transmission capability for about 7 euro: 2 euro for the Arduino and 5 euro for the radio module!

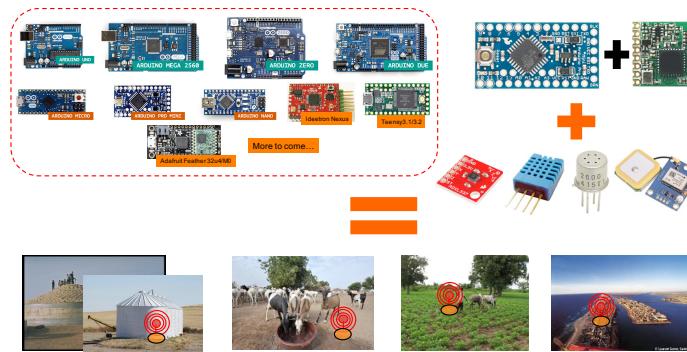


Fig. 2. Generic low-cost IoT hardware

Integration of these generic IoT becomes straightforward and the Arduino Pro Mini is available in the 3.3v & 8MHz version for much lower power con-

sumption, offering the possibility of running for more than a year on 4 AA regular batteries as illustrated in Fig. 3.

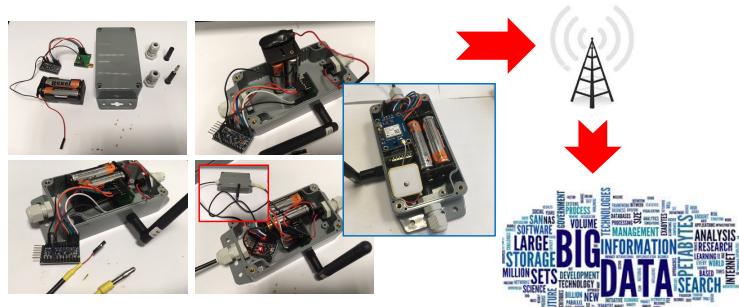


Fig. 3. Easy integration with DIY approach for maximum appropriation

It is expected that this availability of low-cost DIY IoT will create a tremendous uptake of the technology on a large-scale, for a large variety of applications, including those from developing countries as even a limited deployment of IoT devices can have huge impacts.

3 Data analytics with Waziup platform

TODO ►Corentin◀

4 Securing the IoT platform

TODO ►Mehdi: insert content from deliverable◀

5 New market opportunities

TODO ►Abdur: how we detect opportunities/talents through hackathons, local startups, NGOs (such as in Togo), financing etc.◀

6 Related works

In this section, we survey the related works. We first analyse similar works from the literature. Regarding the IoT big data aspect, the research have already been largely instantiated inside Open Source frameworks. We present a selection of the most interesting Open Source contributions.

6.1 Review of literature

IoT in Africa: Their is very little penetration of IoT in Africa, as evidenced in [?] and [?]. The authors of [?] provide a survey, country by country, of the undertaking of IoT. They also document some of the challenges affecting adoption of IoT in the continent. Africa has only 7% of her households on the Internet; this is far behind the worlds figure of 41%. Given this lag in the baseline technology needed to implement Internet of Things, the author of [?] advocate for a technological leap and an African-centric approaches to IoT. Taking the case of a drought early warning and assets tracking systems, the author demonstrates that by innovatively incorporating the realities such as the prevalence of African indigenous knowledge on weather, unreliable communication, low-end mobile phone handsets, among others, a home-grown Internet of Things flavour has higher chance of succeeding. An extensive report from Cisco [?] provides also many insights on the current use and potential of Internet of Things (IoT) technologies in tackling global development challenges, highlighting a number of specific instances where IoT interventions are helping to solve some of the worlds most pressing issues.

A deployment of a Wireless Sensor Network for precise irrigation in Malawi is presented in [?]. For the system to be self-sustained in terms of power, the study used solar photovoltaic and rechargeable batteries to power all electrical devices. The system incorporated a remote monitoring mechanism through a General Packet Radio Service modem to report soil temperature, soil moisture, WSN link performance, and photovoltaic power levels. Irrigation valves were activated to water the field. The paper give insights to develop a robust, fully automated, solar-powered, and low-cost irrigation system to suit the socioeconomic conditions of small scale farmers in developing countries.

The authors of [?] provides a survey of possible IoT applications in South Africa and Zambia. In particular, they identify examples of IoTs to mitigate the agricultural needs of these communities for the domains of crop farming, weather forecasting, wildlife management, forestry, livestock farming, market identification and rural financing.

IoT in Agriculture: There is not a lot of literature on the specific topic of applying IoT in agriculture, and practically none went it comes to rural Africa. In [?], the author presents a work supporting the transition from traditional agriculture to modern agriculture in China. They propose an agriculture intelligent system based on IOT for organic melon and fruit production. A number of new technologies are used, such as RFID and sensors. They monitor temperature, humidity, light and CO_2 around the crops and use a small model on the fruits of growth process. Always in China, [?] uses Internet of things and RFID technologies to realize automatic control production of agriculture.

In [?], the authors perform temperature control in greenhouse using Zigbee. In [?], the authors elaborate a crop growth model. The model is then embedded in their IOT application system. This allows them to make the system more intelligent and adaptive for the facility agriculture. The authors of [?] and [?]

proposes remote agriculture monitoring and process automation. They are both based on gateway infrastructure and wireless connection. The work in [?] shows a semantically enhanced digital agriculture use case built with the OpenIoT platform.

In [?], the authors uses IoT to check electronically on the vital signs of the cattles. Their tool facilitates the day to day management of dairy activities. It also provides forecastings allowing to handle weather related issues, cattle health and emergencies.

IoT is also deployed within the product supply chain, another key area of agriculture. In [?], the authors builds a quantitative trust model to describe the trustworthiness of foods delivered in supply chains. The Internet of Agricultural Things (AIoT), where the technologies of the Internet of Things are widely used in all of the phases in the agriculture industry, is proposed to resolve the food safety problem. In order to provide a common model to describe and transmit the data in agriculture Internet of Things, [?] proposes a specific ontology, while [?] uses a naming service to identify products.

With difference with the literature surveyed in this section, the proposed Waziup platform is a full IoT platform taylored entierly for African need and constraints. In particular, it as application hosting capacities based on the PaaS paradigm, is resilient to disconnections and provides big data capacities.

6.2 Review of Big Data tools

Far from an exhaustive list, this paragraph describes the most used Open Source Big Data tools and compares them in order to give a better understanding of the Big Data ecosystem. Moreover, this review gives an indication on the best tools fitted for WAZIUP platform.

Databases and data warehouses HDFS, developed by Apache, is a distributed, scalable and portable file-system written in Java for the Hadoop Framework⁴. It has been designed for large dataset analysis and by its structure has high fault tolerance. It is the basis upon which everything works in the Hadoop Ecosystem. Build on top of HDFS, Apache HBase is a distributed, non-relational column oriented datastore⁵. HBase is designed to efficiently address random access and fast record lookup. It has the capability to handle extremely large tables of data with low latency. Though, this data storage tool should be used when random and real-time read/write access to data is needed and when many thousands of operation per seconds need to be performed on large datasets (up to petabytes).

Apache Hive⁶ is a data warehouse infrastructure that can manage and query unstructured data as if it were structured. As a full component of Hadoop Ecosystem, it uses MapReduce for execution and HDFS for storage. It has its own language SQL-like (HiveQL) that brings expressiveness to the queries. This

⁴ https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

⁵ <https://hbase.apache.org/>

⁶ <https://hive.apache.org/>

storage mode should be used for SQL-like queries and when higher language than MapReduce is needed. Used by big companies who cant afford to lose data (Apple, Netflix, Spotify), Apache Cassandra⁷ is a column oriented database of structured data. The data are highly available thru column indexes and are automatically replicated thru multiple nodes for fault tolerance. Cassandra has a unique masterless ring design that is easy to setup and to maintain⁸. This tools should be use when losing data is not the critical point and not affordable.

First considered as an outsider, getting rid of traditional table-based relational database, mongoDB⁹ quickly became a must-have tool: a NoSQL, relational, document oriented database. Document are shared in JSON format with dynamic schemas (called BSON) and makes the integration of data sometimes easier. This database is very useful when you need to consume your data in many applications, as many connectors have been developed.

Data publication and subscription Apache Flume¹⁰ is a distributed, reliable and available service for efficiently collecting, aggregating and moving large amounts of streaming event data. Flume should be used if the data is designed for Hadoop as it can move them to HDFS. It has many built-in sources and sinks and can process data in-flight using interceptors, which is useful for data masking or filtering. It is composed of agents and data collectors (and interceptors if needed).

More general purpose, Apache Kafka¹¹ is a high-throughput, distributed, publish-subscribe messaging system. It can replicate events, has low latency and is capable of data partitioning. Kafka is also easily scalable and this tool is very useful when the data is to be consumed by multiple applications. It is composed of producer, consumers and topics. Actually, we might not have to choose between Kafka and Flume, as both can work quite well together. If the workflow design requires streaming data from Kafka to Hadoop, using a Flume agent with Kafka source to read the data makes sense. This association is quite common and is called Flafka¹². If a Data/Context Scenario is developed, we may need to use a context broker. Orion Context Broker¹³ (FIWARE platform) is a publish/subscribe platform that is able to register context elements and manage them through updates and queries. It is possible to subscribe to context information when some conditions occurs (e.g. an interval of time passed or the context elements have changed). Orion is a C++ implementation of the NGSI9/10 REST API binding developed as a part of the FIWARE platform.

⁷ <http://cassandra.apache.org/>

⁸ <http://www.planetcassandra.org/what-is-apache-cassandra/>

⁹ <https://www.mongodb.com/>

¹⁰ <https://flume.apache.org>

¹¹ <http://kafka.apache.org>

¹² <http://blog.cloudera.com/blog/2014/11/flafka-apache-flume-meets-apache-kafka-for-event-processing/>

¹³ <http://catalogue.fiware.org/enablers/publishsubscribe-context-broker-orion-context-broker>

Data processing Obviously, choosing a data processing tool depends mostly on the outcome expected from the data. The most common tool for BigData analysis, and what we probably think at first, is Hadoop MapReduce¹⁴. It has proven its efficiency in many ways and is an incredible tool. But if we want to step a bit aside of Hadoop workflow or if we have specific needs, other tools exists and some are becoming more and more powerful.

But first, lets talk about this milestone Hadoop MapReduce. MapReduce programming model contributed to the amazing progress of BigData processing this past decade. By breaking down the work and recombining it in series of parallelizable operations, it is simple but incredibly efficient and scalable to ten thousands of machines if needed. It can run on inexpensive hardware, lowering the cost of a computing cluster. The latest version of MapReduce is YARN, called also MapReduce 2.0.

If a higher level of programming on top of MapReduce is needed, Apache Pig¹⁵ is the one. Pig has its own language (PigLatin) similar to SQL and works on top of MapReduce. Pig Engine parses, optimizes and automatically executes PigLatin scripts as a series of MapReduce jobs on a Hadoop cluster. Its easy to learn and opens Hadoop to data professionals who may not be software engineers.

First designed to work with HDFS on top of YARN, Apache Spark¹⁶ is a different system for processing data and can work out of Hadoop ecosystem with other data managements systems. It does not work with MapReduce and it can be up to a hundred times faster than MapReduce with its capacity to work in-memory, allowing to keep large working datasets in-memory between jobs, reducing considerably the latency. What makes it more and more attractive to many users worldwide, is its wide range of applications: batch and stream processing (micro-batch processing with 0.5s latency), machine learning (MLib), SQL (with Hive), graph Analytics (graphX). Language supported are Java, Python and Scala.

Demand for stream processing becoming more and more important in Big Data analysis, Apache Flink¹⁷ has been recently developed and is growing very fast. Flink is a streaming dataflow engine that provides data distribution, communication and fault tolerance. It has almost no latency as the data are streamed in real-time (row by row). It runs on YARN and works with its own extended version of MapReduce. Language supported are Java and Scala.

Machine learning Machine learning is the union between statistics and artificial intelligence. It blends AI heuristics with advanced statistical analysis. We let the machine learn about the data, make decisions, and then apply statistics. Algorithms used for this tasks can be grouped in 3 domains of actions: Classification, association and clustering. To choose an algorithm, different parameters must be considered: scalability, robustness, transparency and proportionality. Overlearning (or overfitting) of the model must be carefully checked.

¹⁴ <http://hadoop.apache.org>

¹⁵ <http://pig.apache.org>

¹⁶ <http://spark.apache.org>

¹⁷ <http://flink.apache.org>

Without any math or programming requirement, KNIME¹⁸ is an analytic platform that allow the user to proceed the data in a user-friendly graphical interface. It is a good tool to train your model and evaluate different machine learning algorithms rapidly. If the workflow is already deployed on Hadoop, a machine learning library exists and is called Mahout¹⁹. Spark also has his own machine learning library called MLlib²⁰. H2O²¹ is a software dedicated to machine-learning, which can be deployed on Hadoop or Spark (Flink in development). It has an easy to use Web interface, which makes possible to combine big data analytics easily with machine learning algorithm to train models.

Data visualisation and exploration To visualise the data in real time, Freeboard provides a simple, real-time dashboard, commonly used in IoT world²². There is a direct Orion Fiware connector. To connect with streaming engines, a JSON connector can be used. Design is simple and customisation is not possible, but it is a very good dashboard to visualise easily raw data coming from sensors, before data analysis.

Tableau Public²³ offers a good visualisation and exploration tool on batch data. Tableau is a software where you can upload your analysed data (previously extracted in .csv format). The visualisation tool is very powerful and allow a deep exploration the data. However it is not designed for really Big Data with large datasets and the open Source version of Tableau (Public) does not offer the data streaming capacities (e.g. Spark connectors). Nevertheless, Tableau Public is a highly customisable, user-friendly and intuitive exploration tool for data that have already been processed and analysed.

To visualise data in real-time, after analysis (filtering, aggregating, correlating), one of the best tool is probably Kibana²⁴. It is the visualisation tool coming with ElasticSearch. Elasticsearch is a search server based on Apache Lucene that provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents. It is really designed for real-time analytics, most commonly used with Flink or Spark Streaming.

7 Conclusion

With ICT technologies, Africa can dramatically improve its agricultural productivity by enabling the rapid and cost-effective deployment of advanced and real time monitoring. The immediate effect is to improve coordination and logistics, by reducing time and investment horizons for Research and Development and

¹⁸ <http://www.knime.org>

¹⁹ <http://mahout.apache.org>

²⁰ <http://spark.apache.org/mllib/>

²¹ <http://www.h2o.ai/>

²² <https://freeboard.io/>

²³ <https://public.tableau.com/s/>

²⁴ <https://www.elastic.co/fr/products/kibana>

new product development, and by allowing for the enhanced analysis of historical and ongoing data. With respect to the water sector, such systems can also dramatically improve water use efficiency, allow for the growth of water provider SMEs by providing practical and cost effective new payment, monitoring and management systems. This technology can also offer a new cost-effective alternative for integrated watershed management by networking real-time water quality and flow data. Furthermore, given the fundamental roles which agriculture and water play in the African economic and social development and more generally onto environmental sustainability, WAZIUP can both directly and indirectly bring a much wider range of benefits related to food security, gender equality, poverty reduction and resource use efficiency.

Acknowledgements

This work has been carried out within the European project WAZIUP (H2020-ICT-687607).