Web Semantics: Science, Services and Agents on the World Wide Web 37–38 (2016) 25–35

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

The Graph of Things: A step towards the Live Knowledge Graph of connected things

Danh Le-Phuoc [a,*], Hoan Nguyen Mau Quoc [a], Hung Ngo Quoc [a], Tuan Tran Nhat [a], Manfred Hauswirth [b]

[a] Insight Centre for Data Analytics, National University of Ireland, Galway, Ireland

[b] Institut für Telekommunikationssysteme, Technische Universität Berlin, Berlin, Germany

ARTICLE INFO

Article history:
Received 8 February 2015
Received in revised form
1 January 2016
Accepted 29 February 2016
Available online 11 March 2016

Keywords:
Internet of Things
Graph of Things
Linked Stream Data
Real-time search engine

ABSTRACT

The Internet of Things (IoT) with billions of connected devices has been generating an enormous amount of data every hour. Connecting every data item generated by IoT to the rest of the digital world to turn this data into meaningful actions will create new capabilities, richer experiences, and unprecedented economic opportunities for businesses, individuals, and countries. However, providing an integrated view for exploring and querying such data at real-time is extremely challenging due to its Big Data natures: big volume, fast real-time update and messy data sources. To address this challenge, we provide a unified integrated and live view for heterogeneous IoT data sources using Linked Data, called the Graph of Things (GoT). GoT is backed by a scalable and elastic software stack to deal with billions of records of historical and static datasets in conjunction with millions of triples being fetched and enriched to connect to GoT per hour in real time. GoT makes approximately a half of million stream data sources queryable via a SPARQL endpoint and a continuous query channel that enable us to create a live explorer of GoT (http://graphofthings.org/) with just HTML and Javascript.

1. Introduction

International Data Corporation (IDC) reports that the digital universe will grow by a factor of 300 from 2005 to 2020. Specifically, IDC projects that by 2020 the digital universe will reach 40 zettabytes (ZB), which is 40 trillion GB of data or 5200 GB of data for every person on Earth.[1] The majority of this data will be contributed by billions of devices connected to the Internet of Things (IoT). Connecting every data item generated by IoT to the rest of the digital world to turn this data into meaningful actions will create new capabilities, richer experiences and unprecedented economic opportunity for businesses, individuals and countries. However, deriving trends, patterns, outliers and unanticipated relationships in such enormous amount of dynamic data with unprecedented speed and adaptability is extremely challenging.

Towards unlocking the huge potentials of IoT,[2] deriving insights from dynamic raw IoT data poses various challenges in data integration. As seen on the Web, access to and integration of information from large numbers of heterogeneous IoT streaming sources under diverse ownership and control is a resource-intensive and cumbersome task without a proper support. Such distinct streaming data sources are generated from distributed data acquisition infrastructures of Smart Cities, Social network applications, medical sensors, to name a few. Traditionally, to correlate and analyse them into higher level data products, they need to be transformed, cleaned and consolidated into a large static data warehouse and then made ready for certain types of rudimentary query patterns, e.g. OLAP queries. However, these streaming sources operate on longer time scales on which a wide range of dynamic data feeds are continuously arriving from disparate and uncontrolled sources [1]. Hence, it is much more difficult to maintain a fresh and consistent integrated view for

* Corresponding author.
E-mail address: danh@danhlephuoc.info (D. Le-Phuoc).

1 http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf.

2 http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights/Business%20Technology/Unlocking%20the%20potential%20of%20the%20Internet%20of%20Things/Unlocking_the_potential_of_the_Internet_of_Things_Full_report.ashx.

http://dx.doi.org/10.1016/j.websem.2016.02.003

unlimited discovery and exploration of enormous ever-growing IoT data.

Motivated by such challenges and inspired by the Knowledge Vault [2], we create a Live Knowledge Graph to pave the way towards building a "realtime search engine for the Internet of Things", which we call the Graph of Things (GoT). Similar to the Knowledge Graph used in search engines like Google[3] and Bing,[4] GoT aims at enabling a deeper understanding of the data generated by connected things of the world around us. To tackle the barriers of IoT data interoperability, GoT is represented as Linked Stream Data [3] which employs the RDF data model to use the graph as the unified representation for stream data together with static data. This graph enables a smarter way to discover and explore IoT data under meaningful facts and their relationships.

The effective exploitation of Linked Stream Data from multiple sources requires an infrastructure that supports the intensive effort of enriching, linking and correlating of data streams with very large static data collections, e.g. LinkedGeoData[5] and DBpedia.[6] Moreover, this infrastructure needs to support sophisticated queries, discovery and exploration on increasingly complex data objects representing realistic models of the world. Therefore, we propose a scalable and elastic solution for ingesting, storing, exploring and querying billions of dynamic IoT data points in conjunction with the static datasets from Linked Data Cloud.[7] Our solution also provides an integrated architecture to collect and curate useful RDF-based facts from IoT raw data to create a graph that plays the role as a unified and live view of data objects about "Things". This solution aims to provide a comprehensive software stack with the easy-to-use toolkits to filter, aggregate, enrich and analyse a high throughput of data from multiple disparate live data sources in any data formats. Such data operations are gearing towards identifying simple and complex patterns to visualise business in real-time, to detect urgent situations and to automate immediate actions. To deal with the heterogeneity of dynamic data sources, Linked Stream Middleware [4] of this software stack will transform the data in a variety of data formats and data sources to make it ready for any further processing and high-level analytical operations. Its parallel processing layer of the stack will be able to elastically and dynamically distribute the processing load to the cluster to cope with a large amount of queries and incoming streams while a huge volume of data is already in the store. Our back-end data management system supports the ingestion of million data points per second while it is still able to query live data being indexed to a distributed persistent storage which currently stores billions-triple datasets of historical data as well as static datasets.

The remainder of the article is structured as follows. Section 2 presents the process of building the Graph of Things from both physical world and virtual world. In Section 3, we will share our design choice of building the system and infrastructure to store and query GoT. Section 4 shows some demonstrations and evaluations of our live system and share some lessons learnt during the course of 10 months of deployment. Finally, Section 6 will conclude the article and reveal our future plan for GoT.

3 http://www.google.ie/insidesearch/features/search/knowledge.html.

4 https://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/.

5 http://datahub.io/dataset/linkedgeodata.

6 http://dbpedia.org.

7 http://datahub.io/group/lodcloud.

2. Building a live knowledge graph of connected things

To motivate the creation of the Graph of Things, we start with a real-time search use case as follows. For example, John has just missed his connection flight at Dublin airport, he has few hours to spare in the *sunny weather*. He intends to take the bus to his favourite Sushi restaurant in the city centre, but he finds out there is *no bus today* due to the current *strike of Dublin Bus*. To avoid the current traffic situation in the city centre, he searches for some other Sushi restaurants that *can be reached from the airport in* 10 min *by taxi*. From the list of recommended restaurants, there is a Sushi restaurant *next to Dublin Ferry Port* where an old classmate of John from Liverpool will be *arriving in* 1 h *on a ferry*. Right after finishing the dinner with his friend, he is notified that his next connecting flight *will be delayed for another* 2 h, and he discovers there is an *open-air music show just few blocks from where he is*. From *the live camera feed pointing to the show*, there is a big exciting crowd, he then decides to take a walk there.

In order to provide such live connected information to a user, we have to continuously fuse various IoT stream data sources such as flights, ships, traffic cameras, Tweets and weather sensors into GoT as an integrated view. GoT not only comprises of some data captured by sensors but also involves the context, the meanings and relationships between data objects. In particular, GoT is represented as a big connected RDF graph which enables applications, users and developers to traverse along graph edges (RDF triples) without a restriction on a database schema or fixed connections among data items. The search engine based on this graph can benefit greatly from direct access to these nature relationships to have an intrinsic view of real world events and phenomena. Therefore, the graph can provide smarter search results which might lead the users' curiosities to new relevant topics. In the above use case, the returned data entries to the user like flights, Sushi restaurants and events of interests are graph nodes which are connected to other nodes as the potential data of interest for the user. For instance, the restaurants, the music show events and the ferry have "live" spatial relationships can trigger the interesting correlations according to the user's contexts. Such correlations can be queried by using an extension of SPARQL 1.1 with spatial and temporal filtering conditions for expressing spatial–temporal context of a query. While we defer the technical details on how to express and process such queries, we will present how we model and integrate our IoT data sources summarised in Table 1 in the following sub-sections.

2.1. Collecting facts of physical things

For the physical things equipped with sensors to "observe" facts about them, we use SSN Ontology [5] to capture sensor readings associated with the sensing context including sensor configuration and meaning of what to be measured, etc. To have a richer contextual information, we interlink several datasets from Linked Data Cloud to create meaningful relationships to the sensors, properties and features of interest. For instance, we generate spatial contexts by extracting relevant spatial data from LinkedGeoData, Geonames dataset and other relevant known concepts or entities from DBpedia (which are integrated in YAGO [6]). These relationships play the pivotal roles for correlating sensor readings that share certain contextual links.

The added value of interlinking with the relevant known entities from DBpedia and YAGO is enabling the exploratory search for users who are unsure about their goals in the first place. For instance, a user looks for relevant real-time sensor data of a *blizzard* happening in his/her city. The easiest way is to start a keyword search, *blizzard*, which can returns relevant properties of the YAGO entity, <*wordnet_blizzard*_111509570>. By following this entity,