

Text Analytics and Linked Data Management As-a-Service with S4

Marin Dimitrov, Alex Simov, Yavor Petkov

Ontotext AD, Bulgaria
{first.last}@ontotext.com

Abstract. One of the limiting factors for the wider adoption of Semantic Technology at present is the complexity and cost of existing enterprise solutions for text analytics and Linked Data management. Startups and mid-size businesses often have only limited resources to evaluate and prototype with novel approaches for semantic data management. The Self-Service Semantic Suite (S4) provides an integrated platform for on-demand, cloud-based text analytics and Linked Data management. With S4 the companies in the early stages of evaluating and adopting Semantic Technology have the ability to easily and quickly apply a full suite of semantic data management and text analytics within applications in various domains.

Keywords: text analytics, cloud computing, software-as-a-service, database-as-a-service, linked data, knowledge graphs, semantic web

1 Introduction

Semantic Technologies provide a family of novel approaches for data integration, discovery and analytics. Various successful applications of Semantic Technology were introduced by big enterprises, including Fortune 500 companies, within the last three years. At the same time, the wider adoption of these technologies is still slower than expected and Gartner usually positions them in the early phases of its regular technology *hype cycle*¹ analysis.

An additional limiting factor for the wider adoption of Semantic Technology at present is the complexity and cost of existing enterprise solutions for text analytics and Linked Data management. Startups and mid-size businesses have very limited resources to evaluate and prototype with novel technologies and approaches for semantic data management: on premise hardware and licensing costs create additional barriers to entry. Enterprise organizations often have complex and slow procurement processes and they are losing valuable innovation cycles by not having easy access to this technology.

¹ <http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp>

2 Goals and Use Cases

The Self-Service Semantic Suite² (S4) provides a platform for on-demand text analytics and Linked Data management. With S4 the companies which are in the early stages of evaluating or adopting Semantic Technology will be able to instantly access a full suite of semantic data management and text analytics capabilities, without the need for complex planning, investments and operations.

The main benefits of using on-demand text analytics and Linked Data management capabilities can be summarised as follows:

- *Shorter time-to-market* – companies who are currently experimenting with Semantic Technology and fall into the groups of *technology enthusiasts* and *visionaries*, based on their attitudes towards adopting an emerging technology [1], can benefit from capabilities for semantic data management available from the “get-go” and without the need for complex on-boarding, integration and customization. Such organisations will be able to deliver new prototypes faster and at a lower cost, while looking for a successful breakthrough.
- *Risk reduction* – companies who fall in the group of *pragmatists* regarding technology innovation, can benefit from a low-risk option for experimenting with Semantic Technology, without the need to commit to license and hardware investments and deal with provisioning and operations overheads. By using a platform for on-demand semantic data management, such companies can thoroughly evaluate the Semantic Technology maturity and expected ROI potential within their vertical.
- *Cost optimisations* – companies who have already evaluated the Semantic Technology potential ROI and are committed to its long term adoption can often realise cost reductions by switching from a traditional model of on-premise software deployment towards a more flexible on-demand and pay-per-use cost model.

3 The Self-Service Semantic Suite

S4 provides a set of capabilities covering key aspects of the semantic data management lifecycle:

- Reliable access to central *Linked Open Datasets* such as DBpedia, Freebase, GeoNames, MusicBrainz and WordNet.
- A self-managed and a fully-managed scalable *RDF database as-a-service* in the Cloud for private RDF knowledge graphs.
- Various *text analytics services* for news, biomedical documents and social media content (Twitter).
- Web based *data-driven portals for RDF* search and exploration.

² <http://s4.ontotext.com/>

3.1 Linked Open Data access

Among the main challenges related to the application of Linked Data in enterprise contexts are the performance and availability problems associated with most of the public LOD endpoints at present [2]. S4 provides a reliable access to key datasets from the LOD cloud via the FactForge³ large-scale semantic data warehouse [3]. More than 5 billion LOD triples, describing 500 million entities from integrated and aligned datasets – such as DBpedia, Freebase, GeoNames, and MusicBrainz – are available to S4 developers. In the near future S4 will be extended to provide access to other key Open Data and Linked Data sets as well.

3.2 RDF Database As-a-Service

S4 provides an RDF database-as-a-service based on one of the leading RDF databases: GraphDB⁴ [4]. The database-as-a-service capability of S4 is available in two flavours: a *self-managed* cloud database, where the user is in full control of operational aspects – such as availability, performance tuning, backups and restores – and a *fully managed* cloud database, where the S4 platform takes care of all aspects related to database administration, provisioning and operations.

The self-managed database provides on-demand private database servers (single tenant model) for organizations which need only the occasional, yet high-performance and reliable access to private RDF datasets, in cases where an on-premise software and hardware deployment would not be cost optimal.

The fully managed RDF database provides a 24/7 access to private RDF datasets and SPARQL endpoints within a multi-tenant model. All operational aspects such as security, availability, monitoring and backups are handled by S4 on behalf of the users. The fully managed database has a container-based architecture (with Docker⁵ technology) for improved security isolation and resource utilisation control of the different database instances hosted within the same virtual machine.

3.3 Text Analytics As-a-Service

S4 provides various services for real-time text analytics over unstructured content:

- *News Analytics* – the service performs information extraction, disambiguation and entity linking to DBpedia, Freebase and GeoNames. The text analysis process is a combination of rule-based and machine learning techniques [5].
- *News Classifier* – the service performs categorisation of news articles according to the 17 top-level categories of the IPTC Subject Reference System [6].

³ <http://factforge.net/>

⁴ <http://www.ontotext.com/products/ontotext-graphdb/>

⁵ <https://www.docker.com/>

- *Biomedical Analytics* – the service can recognize more than 130 biomedical entity types [7] and semantically link them to a large-scale biomedical LOD knowledge base (LinkedLifeData⁶).
- *Twitter Analytics* – the service is based on the TwitIE open source microblog analysis pipeline [8] and it performs named entity recognition of various classes of entities as well as normalisation of most common abbreviations frequently found in tweets.

3.4 Data-driven Portals

The data-driven portals provide a simple way to query and explore the RDF data stored in S4. The portals provide means for SPARQL querying, RDF data exploration and navigation and faceted search. The data-driven portals are based on the GraphDB Workbench technology [10], though in the future 3rd party tools with more powerful capabilities for RDF data exploration and visualisation will be incorporated into S4.

3.5 Public Cloud Platform

S4 is deployed on a public AWS⁷ cloud platform and it utilizes various cloud infrastructure services such as:

- *distributed storage* via Simple Storage Service, Elastic Block Storage, DynamoDB and Glacier
- *computing* via Elastic Compute Cloud, Auto Scaling, Elastic Load Balancer and AWS Lambda
- *application integration* via Simple Queue Service, Simple Email Service and Simple Notification Service

S4 is designed for a multi-datacenter deployment for improved resilience and availability. This is an important design decision for such platforms, since even though public cloud outages are quite rare and usually short, business continuity guarantees are crucial for the adoption of an as-a-service based technical solution.

3.6 Architecture

The architecture of S4 is based on the best practices and design patterns for scalable cloud AWS architectures [11]. The initial architecture of S4 was based on our previous work on the AnnoMarket platform [9], though the current architecture has been extended in order to accommodate the RDF database-as-a-service and improved with various features related to scalability, reliability and security.

S4 follows the principles of micro-service architectures and it is comprised of the following main components and layers:

⁶ <http://linkedlifedata.com/>

⁷ <http://aws.amazon.com/>

- *Load balancer* – the entry point to all S4 services is the load balanced of the AWS platform, which will route incoming requests to one of the available frontend nodes. The load balancer can distribute requests even between instances in different datacenters.
- *Frontend nodes* – the frontend nodes host various micro-services such as: user management, text analytics frontend, as well as the front-end services for the LOD server and the RDF database-as-a-service layer. All instances host the same set of stateless front-end services and the frontend layer is automatically scaled up or down (new instances added or removed) based on the current system load.
- *Text analytics nodes* – these nodes are responsible for processing the text documents sent for analysis to S4. They host the different text analytics pipelines for news, biomedical documents and social media. This layer is also automatically scaled up or down based on the current system load.
- *Database nodes* – this layer contain nodes running multiple instances of the GraphDB database (packaged as Docker containers). Each user has its own database instance (container) and it cannot interfere with the database instance or with the data of the other users of the platform. The data is hosted on Network-attached storage volumes (EBS) and each user/database has its own private EBS volume. Additional OS level security ensures the proper data isolation and access control. Unlike the other layers of the system, each virtual machine in this layer hosts only a subset of all the database containers, e.g. database containers are not replicated across backend servers. Future versions of S4 will introduce container replication as well for the purpose of improved throughput, so that read-only queries can be distributed among multiple servers hosting same database replica.
- *Linked Data server* – currently the LOD data available through S4 is hosted on the FactForge semantic data warehouse.
- *Integration services* – a distributed queue and a distributed push messaging service are used for loose coupling between the various frontend and backend nodes on the platform. For example, the requests for text processing are first handled by one of the frontend nodes, which puts a processing request in the distributed queue and one of the available text analytics backend nodes will pull the request, process it, and send the result back to the frontend node. This way, the frontend and the backend layers are not aware of their size and topology and they can be scaled up or down independently.
- *Distributed storage* – S3 is used for transient storage of text documents to be processed, while all persistent data is stored on the Network-attached Storage (EBS). Logging data, user data as well as various configuration metadata is stored in a distributed NoSQL database (DynamoDB).
- *Management services* – various management services are available on the S4 platform: logging, reporting, account management, quota management, and billing.
- *Monitoring services* – the AWS cloud provides various metrics for monitoring the service performance. S4 utilises these metrics in order to provide optimal performance and scalability of the platform. The different layers of the platform can be automatically scaled up (to increase system performance) or down (to decrease operational costs) in response to the current system load and utilisation.

3.7 Add-ons

In order to assist developers with using the various S4 services, the platform provides different plugins and add-ons to 3rd party tools:

- An add-on for the General Architecture for Text Engineering⁸ (GATE) platform, which makes it easy for GATE developers to embed S4 text analytics services in complex text processing workflows and applications. A similar plugin, as well as an SDK, is available for the Unstructured Information Management Architecture⁹ (UIMA) text analytics platform.
- A Firefox and Chrome browser plugins which allow web page snippets to be quickly annotated with S4 text analytics services directly from the browsers.
- Java and C# SDKs that provide developers with easy access to the S4 services in their programming language of choice.

4 Lessons Learned

The lessons learned and best practices that emerged during the design, implementation and operation of the S4 platform can be summarised as follows:

- *“Cost-aware” architecture* – the term was initially coined by the Amazon CTO Werner Vogels to describe cloud architectures where operational cost increases are well aligned with the revenue growth. In the case of S4, the architecture allows for elastic scaling up or down of the number of computing and storage nodes on the platform, based on number of text documents waiting for processing, the number of triples stored in the system, or the Linked Data server or RDF database-as-a-service query load – the actual revenue dimensions for S4. The platform is designed so that it can quickly adapt to increased usage by dynamically provisioning extra capacity, and then scale down when the usage decreases in order to optimise the operational costs.
- *Extensive system benchmarking* – S4 is utilising more than a dozen of the AWS cloud platform services, which are available in multiple configurations and pricing plans. Optimising the architecture in terms of cost requires an extensive benchmarking with real-world test cases and data volumes, so that the optimal price/performance configuration for each component of the architecture can be identified. The benchmarking process should be an important part of the architecture and design, because any as-a-service platform should be able to adapt and respond even to rapid changes and fluctuations of its load and utilisation.
- *Microservice and cloud-native architecture* – designing a platform that optimally utilises the underlying Cloud platform is significantly more challenging than just deploying an existing software platform or an application in the Cloud. Nonetheless, a Cloud-native and microservice based architecture makes it possible for vari-

⁸ <http://gate.ac.uk/>

⁹ <https://uima.apache.org/>

ous the S4 components and services to be continuously improved and re-deployed multiple times per week without any service interruptions and system-wide downtime.

- *Resilient design* – S4 follows the best practices for resilient design, since the complexity of a distributed Cloud infrastructure makes inevitable failures on various levels of severity – reduced performance of a Cloud service, partial or full service failure, partial or full datacentre failure, or even failures affecting multiple datacentres at the same time. S4 is designed in a way that minimises the impact of failures in the underlying Cloud infrastructure and improves the business continuity of S4 services.

5 Related Work

Several companies currently offer a text analytics as-a-service capability: OpenCalais, Alechemy, OpenAmplify, Semantria, TextWise, Saplo, Bitext, etc. Some RDF database vendors also provide options for self-managed (OpenLink) or fully managed (Dydra) RDF databases in the cloud. Several public LOD endpoints are provided and maintained by different organisations.

The main differentiation of the S4 platform is that it provides an *integrated suite* of key semantic data management capabilities: 1) reliable access to central LOD sets and knowledge graphs; 2) scalable RDF databases in the cloud, available in a self-managed or fully managed way; 3) text analytics components for news, life sciences and social media; 4) data-driven portals for data discovery & exploration. Such capabilities allow for content from unstructured data sources to be analysed, enriched and interlinked into knowledge graphs, so that semantic search and discovery can be performed.

S4 will be additionally extended with services covering the full lifecycle of semantic data management and analytics.

Acknowledgements.

Some of the work related to S4 is partially funded by the European Commission under the 7th Framework Programme, project DaPaaS¹⁰ (No. 610988)

References

1. G. Moore. *Crossing the Chasm: Marketing and Selling Disruptive Products to Mainstream Customers* (3rd ed). HaperBusiness. 2014.
2. C. Buil-Aranda, A. Hogan, J. Umbrich, and P. Vandenbussche. SPARQL Web-Querying Infrastructure: Ready for Action? In *Proceedings of the 12th International Semantic Web Conference*, Sydney, Australia. 2013

¹⁰ <http://project.dapaas.eu/>

3. M. Damova, K. Simov, Z. Tashev, and A. Kiryakov. FactForge: Data Service or Diversity through Inferred Knowledge over LOD. In *Proceedings of AIMS'A'2012*. Varna, Bulgaria. 2012
4. B. Bishop, A. Kiryakov, D. Ognyanoff, I. Peikov, Z. Tashev, and R. Velkov, OWLIM: A family of scalable semantic repositories. In *Semantic Web Journal*, vol 2, number 1. 2011
5. G. Georgiev, B. Popov, P. Osenova, and M. Dimitrov. Adaptive Semantic Publishing. In *Workshop of Semantic Web Enterprise Adoption and Best Practice (WaSABi)* at ISWC 2013. Sydney, Australia, CEUR WS Vol-1106. 2013
6. IPTC. Subject Reference System Guidelines. Available at http://www.iptc.org/std/NewsCodes/0.0/documentation/SRS-doc-Guidelines_3.pdf. 2003.
7. G. Georgiev, K. Pentchev, A. Avramov, T. Primov, and V. Momtchev. Scalable Interlinking of Bio-Medical Entities and Scientific Literature in Linked Life Data. In proceedings of CALBC workshop. 2011
8. K. Bontcheva, L. Derczynski, A. Funk, M.A. Greenwood, D. Maynard, and N. Aswani. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*. 2013
9. M. Dimitrov, H. Cunningham, I. Roberts, P. Kostov, A. Simov, P. Rigaux, and H. Lippell. AnnoMarket – Multilingual Text Analytics at Scale on the Cloud. In European Semantic Web Conference (ESWC) Poster & Demo proceedings, Hersonissos, Greece. 2014
10. Ontotext. GraphDB Workbench Users Guide. Available at <http://owlim.ontotext.com/display/GraphDB6/GraphDB-Workbench>. 2014
11. Amazon Web Services. AWS Reference Architectures. Available at <http://aws.amazon.com/architecture/>. 2014