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 a wider adoption of Semantic Technology at present is the complexity and cost for deploying and licensing existing enterprise solutions for text analytics and Linked Data management. Startups and mid-size businesses often have only limited resources to evaluate and prototype such novel approaches for semantic data management. The Self-Service Semantic Suite (S4) provides an integrated platform for cloud-based text analytics and linked data management as-a-service. With S4 companies who are in the early stage of evaluating and adopting Semantic Technology have the ability to easily, quickly and at a low cost apply a full suite of semantic data management and text analytics within applications in various domains.

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

1 Introduction

Semantic Technologies provide a family of novel approaches for data analysis, integration and discovery. Several high-impact 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 a wider adoption of Semantic Technology at present is the complexity and cost for deploying and licensing 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 such novel approaches for semantic data management: on premise hardware and software 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 Suite[2] (S4) aims at reducing the cost and complexity of Semantic Technology adoption by providing an integrated platform for cloud-based text analytics and linked data management as-a-service. With S4 the companies which are in the early stages of evaluating and adopting Semantic Technology will have the ability to easily, quickly and at a low cost apply a full suite of semantic data management and text analytics within applications suited for various domains, without the need for complex planning, budgeting, investment, provisioning and operations.

Given the above challenges, there are several potential use cases for text analytics and Linked Data management as-a-service similar to S4:

- *Reducing time-to-market* – companies who experiment with Semantic technology and who fall into the groups of *technology enthusiasts* and *visionaries*, based on their openness towards adopting a technology innovation [1], need capabilities for semantic data management and text analytics that are available from the "get-go" and do not require complex on-boarding, integration and customization, so that the organisations can deliver new product and service prototypes at a rapid rate, while looking for the successful breakthrough / minimum viable product.
- *Reducing risk* – companies who fall in the group of *pragmatists* when it comes to innovation adoption rate, benefit from a low-cost and low-risk option for experimenting and adopting Semantic Technology, without the need to commit to license purchases, hardware provisioning and complex procurement procedures. By using a platform for semantic data management and text analytics as-a-service, these companies can thoroughly evaluate the Semantic Technology maturity, reliability, performance and ROI potential before committing to adopting it.
- *Optimising costs* – companies who have already evaluated the Semantic Technology potential ROI and are committed to its long term adoption and use can often reap cost reductions by switching from a traditional model of on premise licensed software deployment towards a public/hybrid cloud deployment with pay-per-use cost model.

3 The Self-Service Semantic Suite

S4 provides a self-service and on-demand set of components for low cost semantic data management and text analytics, covering key aspects of the data management lifecycle:

- On-demand, fast and 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 available as-a-service* in the Cloud, for private RDF knowledge graphs

[2] http://s4.ontotext.com/

- various *text analytics services* for news, biomedical documents and social media content (Twitter) that extract valuable insight from unstructured content
- web based *data-driven portals for RDF* search and exploration.

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 many public LOD endpoints [2]. S4 provides a fast, reliable and metered access to key datasets from the LOD cloud via the FactForge[3] large scale semantic data warehouse [3]. More than 5 billion LOD triples are available to S4 developers via FactForge, from integrated and aligned datasets such as DBpedia, Freebase, GeoNames, and MusicBrainz as well as ontologies and vocabularies like DC, SKOS, FOAF and PROTON[4].

In the near future S4 will be extended to provide a reliable access to other key Open Data and Linked Data sets.

3.2 RDF Database As-a-Service

S4 provides an RDF database-as-a-service capability based on one of the leading enterprise RDF databases: GraphDB[5] (formerly OWLIM) [4].

The cloud database infrastructure 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, backup and restores, etc., and a *fully managed* cloud database where the S4 platform takes care of database administration and operations tasks.

The self-managed database in the cloud provides an on-demand and private database server (single tenant model) with a pay-per-use (per hour of use) pricing. It provides a cost efficient enterprise RDF database to organizations which need only 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 efficient.

The fully managed RDF database in the cloud provides a low-cost and pay-per-use (per number of triples stored and queries executed) 24/7 access to private RDF datasets and SPARQL endpoints within a multi-tenant model. Operational aspects such as security, availability, monitoring and backups are fully handled by S4 on behalf of the users. The security isolation and resource utilisation control of the different database instances hosted within the same virtual machine in the Cloud is achieved by employing a container-based architecture with the Docker[6] technology.

3.3 Text Analytics As-a-Service

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

---

[3] http://factforge.net/
[4] http://dev.ontotext.com/proton-ontology
[5] http://www.ontotext.com/products/ontotext-graphdb/
[6] https://www.docker.com/

- *News* – the news analytics service performs information extraction and entity linking to DBpedia, Freebase and GeoNames. The text analysis process is a combination of rule-based and machine learning techniques. The service applies word sense disambiguation techniques and attaches a unique URI to each extracted entity or relation from the text [5].
- *Biomedical documents* – the biomedical text analytics service [6] can recognize more than 100 biomedical entity types and semantically link them to a large scale biomedical LOD knowledge base (LinkedLifeData[7]).
- *Twitter* – the Twitter analytics service is based on TwitIE [7] 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 but efficient way to query and explore the RDF data stored in the fully managed cloud database of S4. The portals provide means for SPARQL querying, RDF data exploration and navigation, faceted search, export and import of data, as well as simple administrative tasks such as security and access control specification for the data stored in the customer's private RDF database. The data-driven portals are based on the GraphDB Workbench technology [9], though in the future more powerful means for RDF data exploration and visualisation may be incorporated into S4.

3.5 Public Cloud Platform

S4 is currently deployed on a public AWS[8] cloud platform and it utilizes various cloud infrastructure services such as:

- *distributed storage* via Simple Storage Service, Elastic Block Storage and DynamoDB
- *computing* and *scalability* via Elastic Compute Cloud, Auto Scaling and Elastic Load Balancer
- *application integration* via Simple Queue Service and Simple Email Service
- *system monitoring* via CloudWatch and CloudTrail

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.

---

[7] http://linkedlifedata.com/

[8] http://aws.amazon.com/

3.6 Architecture

The architecture of S4 is based on best practices and design patterns for scalable cloud architectures [10]. The initial design and architecture of S4 was based on our previous work on the AnnoMarket platform [8], though the current architecture has been extended 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 layers:

- *Load balancer* – the entry point to *all* S4 services is the load balanced of AWS which will route incoming requests to one of the available frontend instances. The load balancer can distribute requests even between instances in different datacenters, as long these centers are co-located (within the same *availability zone*, in AWS terminology)
- *Frontend instances* – the frontend instances 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-as-service instances in the backend. All virtual machines host the same set of 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 backend* – these instances 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 as-a-service backend* – this layer contain virtual machines that host running GraphDB instances (packaged as Docker containers). Each user has its own database instance (container) and cannot interfere with neither with the database instance nor with the data of the other users. The data is hosted on a Network Attached Storage volumes (EBS) and each user/database has its own private EBS volume. Additional OS level security ensures the proper isolation of user's data. Unlike the other layers of the system, each virtual machine in this layer hosts a subset of all the database containers, e.g. at present the containers are not replicated. Future versions of S4 will introduce container replication as well, so that SPARQL queries can be distributed among multiple machines serving the same database for the purpose of improved query throughput.
- *Linked Data server* – currently the LOD data available through S4 is hosted on the FactForge semantic data warehouse.
- *Integration services* – a distributed queue is used for loose coupling between frontend and backend components. For example the requests for text processing (sent either via the API or via plugins) are first handled by the frontend, which puts a processing request in the queue and one of the running text analytics backend machines (based on its current load) will pull the document, process it, and send the result back to the frontend instance. This way the frontend and the backend are not aware of their size or topology and they can be scaled up/down independently.

- *Distributed storage* – S3 is used for temporary/transient storage, 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, billing module, payments module, etc. The management services are not subject to automatic scaling up/down due to their low utilisation – a single virtual machine is sufficient for performing these tasks.
- *Monitoring services* – the AWS cloud provides various metrics and means for monitoring service performance. S4 utilises these metrics in order to provide good performance and scalability of the platform and various layers of the platform can be automatically scaled up (to increase system performance) or down (to decrease operational costs).

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:

- A Processing Resource plugin for the General Architecture for Text Engineering[9] (GATE) platform [11] which makes it easy for GATE developers to embed S4 text analytics services in arbitrary text processing pipelines and applications.
- A Firefox plugin which allows web page snippets to be annotated with S4 text analytics services.
- A Java API that provides developers with easy access to the S4 services.

In the future additional S4 add-ons will target more tools such as WordPress, more browsers, and ontology editors as well as APIs for more programming languages.

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 Werber 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 scaling up/down based on number of documents waiting for processing, number of triples stored in the system as well as SPARQL query load – the potential 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.

[9] http://gate.ac.uk/

- *Extensive system benchmarking* – S4 is utilising more than a dozen of cloud 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, so that the optimal price/performance balance for each component of the architecture can be identified. The benchmarking process should be an important part of the architecture and design, because an as-a-service platform should be able to quickly adapt even to big changes in the system load.
- *Micro-service and cloud native architecture* – designing a platform that optimally utilises the underlying Cloud platform is significantly more challenging that just adapting and deploying an existing system or an application on the Cloud. Nonetheless a cloud native and micro-service based architecture makes it possible for various S4 components and services to be continuously improved and re-deployed multiple times per week without any service interruptions and 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 component, partial Cloud service failure, full service failure, partial/full datacentre failure, or even partial/full region failure. 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, 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 as-a-service, anytime, anywhere, within a pay-per-use model: 1) reliable access to central LOD sets; 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.

S4 will be additionally extended to cover 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 DaPaaS[10] (No. 610988)

[10] http://project.dapaas.eu/

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
3. M. Damova, K. Simov, Z. Tashev, and A. Kiryakov. FactForge: Data Service or Diversity through Inferred Knowledge over LOD. In *Proceedings of AIMSA'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. 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
7. 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
8. 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
9. Ontotext. GraphDB Workbench Users Guide. Available at http://owlim.ontotext.com/display/GraphDB6/GraphDB-Workbench. 2014
10. Amazon Web Services. AWS Reference Architectures. Available at http://aws.amazon.com/architecture/. 2014
11. H. Cunningham, et al. Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science. ISBN 0956599311. 15 April 2011.