Contents lists available at ScienceDirect

Future Generation Computer Systems

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

SWAN: A service for interactive analysis in the cloud

Danilo Piparo [*], Enric Tejedor, Pere Mato, Luca Mascetti, Jakub Moscicki, Massimo Lamanna

CERN, CH-1211 Geneva 23, Switzerland

HIGHLIGHTS

- A new service for web-based data analysis in the cloud is proposed: SWAN.
- SWAN combines CERN IT services with modern technologies of distributed computing.
- Synchronised cloud storage is a cornerstone of a data analysis service.
- SWAN users can synchronise their cloud storage space in their local machines.
- A cloud directory is the sharing unit of scientific results: code, text and data.

ARTICLE INFO

Article history:
Received 1 April 2016
Received in revised form
27 October 2016
Accepted 30 November 2016
Available online xxxx

Keywords:
File sharing
Storage synchronisation
Cloud
Analysis
Mining
Notebook

ABSTRACT

SWAN (Service for Web based ANalysis) is a platform to perform interactive data analysis in the cloud. SWAN allows users to write and run their data analyses with only a web browser, leveraging on the widely-adopted Jupyter notebook interface. The user code, executions and data live entirely in the cloud. SWAN makes it easier to produce and share results and scientific code, access scientific software, produce tutorials and demonstrations as well as preserve analyses. Furthermore, it is also a powerful tool for non-scientific data analytics.

This paper describes how a pilot of the SWAN service was implemented and deployed at CERN. Its backend combines state-of-the-art software technologies with a set of existing IT services such as user authentication, virtual computing infrastructure, mass storage, file synchronisation and sharing, specialised clusters and batch systems.

The added value of this combination of services is discussed, with special focus on the opportunities offered by the CERNBox service and its massive storage backend, EOS. In particular, it is described how a cloud-based analysis model benefits from synchronised storage and sharing capabilities.

1. Introduction

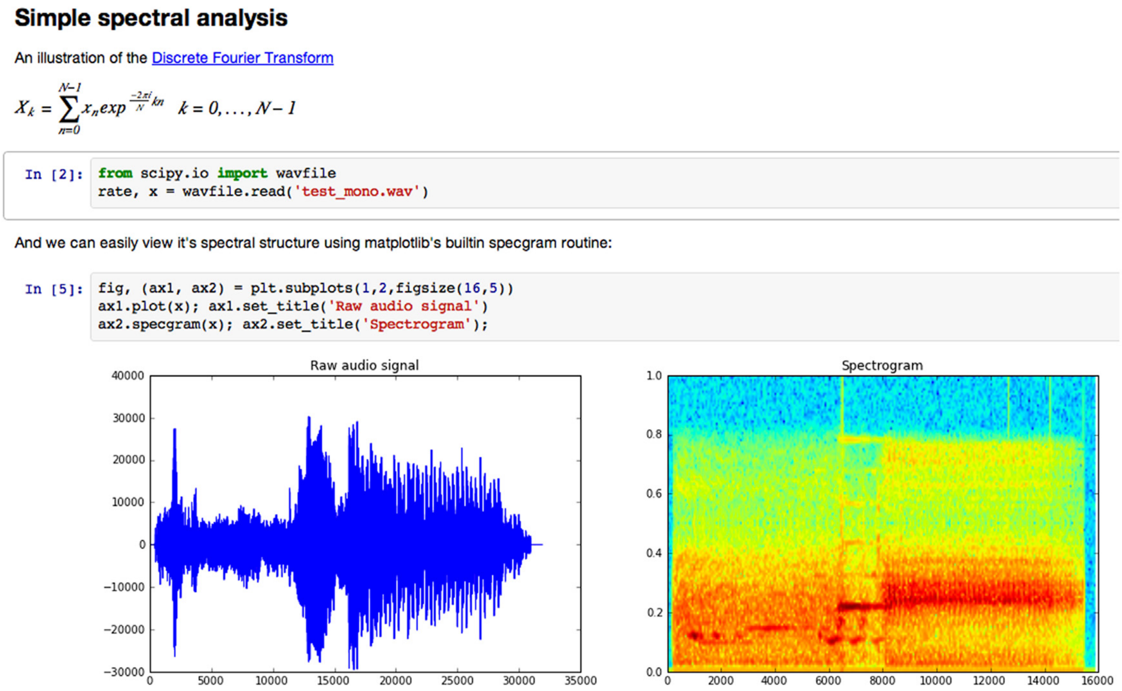
For several years, High Energy Physics (HEP) has been facing unprecedented challenges in data storage, processing and analysis. As an example, the Large Hadron Collider (LHC) experiments at CERN [1] generate about 40 terabytes/s of raw data, which, after processing and filtering, results in tens of petabytes per year. During the last decade, the Worldwide LHC Computing Grid (WLCG) [2] has provided the infrastructure to store, distribute and analyse all this data. Scientists from around the world submit their jobs to the WLCG grid resources to execute their analyses on a daily basis.

Nevertheless, HEP is not the only community that has to confront with the big data challenge. Other examples in science include astronomy [3] and bioinformatics [4]. Industry is clearly leading the way in the field, especially big companies like Google, Amazon or Facebook, which mine customers' data for sales and marketing purposes [5]. Smaller-size organisations also have the means to collect and analyse fairly big amounts of data, mainly thanks to open source tools like Hadoop [6].

Among the directions explored by those communities, there is a noticeable trend towards *web-based interactive analysis*, where the user interacts with an on-line service by means of a web-browser [7–10]. This "software as a service" provisioning model allows users to focus on the solution of a problem in question rather than on installation, configuration and operational matters. Furthermore, such services are often backed up by computing and data resources that are hosted "in the cloud".

[*] Corresponding author.
E-mail address: danilo.piparo@cern.ch (D. Piparo).

http://dx.doi.org/10.1016/j.future.2016.11.035
0167-739X/© 2016 Elsevier B.V. All rights reserved.



Simple spectral analysis

An illustration of the Discrete Fourier Transform

$$X_k = \sum_{n=0}^{N-1} x_n exp^{\frac{-2\pi i}{N} kn} \quad k = 0, \dots, N-1$$

```
In [2]: from scipy.io import wavfile
        rate, x = wavfile.read('test_mono.wav')
```

And we can easily view it's spectral structure using matplotlib's builtin specgram routine:

```
In [5]: fig, (ax1, ax2) = plt.subplots(1,2,figsize(16,5))
        ax1.plot(x); ax1.set_title('Raw audio signal')
        ax2.specgram(x); ax2.set_title('Spectrogram');
```

Fig. 1. An example of a Jupyter notebook [14]: text, formulae, code and images are combined in a computational narrative. Interactive JavaScript based widgets can also be used to provide increased interactivity and multimedia approach, e.g. for playing sounds or display videos [15].

These circumstances led to rethinking the data analysis models at CERN, more precisely in two ways: (i) how physicists could benefit from a service for interactive data analysis in the cloud, with only a web browser, and (ii) how state-of-the-art tools and existing CERN technologies could be combined to implement such a service, making it possible to access computing and storage resources transparently and on demand.

In particular, the cloud storage component plays a key role in the service and should fulfil three basic requirements: be the reference backend for both end-user and experiment data, be responsive enough to provide a good interactive user experience and provide easy means for scientists to synchronise their local workspaces and share their analyses.

In that sense, this paper presents the Service for Web based data ANalysis (SWAN), a cloud-based and interactive data analysis platform accessible via a web interface. SWAN boosts the productivity of scientists and engineers by allowing them to focus solely on the solution of their problems without investing resources in the creation, configuration and maintenance of software and hardware environments. Moreover, it facilitates the sharing of results and code, the access to scientific software, the achievement of reproducible results, the creation of tutorials, demonstrations for outreach and teaching, providing also many necessary elements for the preservation of data analysis procedures.

This paper is structured as follows. Section 2 introduces the interface that was chosen for SWAN, mainly based on the Jupyter [11] notebook platform for interactive data analysis. Section 3 describes the implementation of the service backend, i.e. how production-grade CERN IT technologies can be orchestrated in combination with other cutting-edge tools to build a distributed computing infrastructure. In Section 4 the role of the storage component in the service is characterised in detail. Section 5 is dedicated to the review of a series of use-case categories which are targeted by SWAN. Section 6 illustrates the main features of other work related to SWAN. Finally, Section 7 discusses the conclusions and Section 8 future work.

2. The notebook interface for interactive analysis

A common approach for interactive data analysis is to combine code, text, plots and rich media in the same document, known as *notebook*. Notebooks are divided in cells where the user can type code, execute it and see the results inline. In short, they can be thought of as an interactive programming shell running in a web browser.

There exist several notebook flavours [8,12,13], although one of them has been particularly successful: the Jupyter open source project [11]. Jupyter notebooks are an agile tool for both exploratory computation and data mining, and provide a platform to support reproducible research, since all inputs and outputs may be stored in a one-to-one way in the same document. Fig. 1 illustrates an example of a Jupyter notebook document where markdown text, formulae, code and plots are combined.

On the other hand, Jupyter is not restricted to a particular programming language, but instead it allows to plug in language extensions known as kernels. At the time of writing, more than forty programming languages are supported. Moreover, Jupyter can accommodate various ecosystems of tools for data analysis, e.g. R [16], Numerical Python [17] or Pandas [18].

The aforementioned attractive features of the Jupyter notebooks motivated their choice as the main interface of the SWAN service. Thus, users are able to produce their analyses in the form of notebooks by using only a web browser. The execution of the notebook cells, as well as the management of their associated data, happens seamlessly and transparently in the cloud. In order to support this cloud execution model of user notebooks, a service backend is needed; the details of such implementation are discussed next in Section 3.

3. SWAN and the portfolio of CERN services

The web-based interface of SWAN, based on Jupyter notebooks, is powered by a service backend that manages the execution of the notebooks on behalf of the users. Fig. 2 depicts the design of the SWAN backend, which strongly relies on a portfolio of already

Please cite this article in press as: D. Piparo, et al., SWAN: A service for interactive analysis in the cloud, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.11.035