Geophysical Research Abstracts
Vol. 16, EGU2014-14132-2, 2014
EGU General Assembly 2014

A lightweight messaging-based distributed processing and workflow execution framework for real-time and big data analysis

Shaban Laban (1) and Aly El-Desouky (2)

(1) CTBTO Preparatory Commission, International Data Centre, Vienna, Austria (shaban.laban@ctbto.org), (2) Computer and Systems Department, Faculty of Engineering, Mansoura University, Egypt.

To achieve a rapid, simple and reliable parallel processing of different types of tasks and big data processing on any compute cluster, a lightweight messaging-based distributed applications processing and workflow execution framework model is proposed. The framework is based on Apache ActiveMQ and Simple (or Streaming) Text Oriented Message Protocol (STOMP). ActiveMQ , a popular and powerful open source persistence messaging and integration patterns server with scheduler capabilities, acts as a message broker in the framework. STOMP provides an interoperable wire format that allows framework programs to talk and interact between each other and ActiveMQ easily. In order to efficiently use the message broker a unified message and topic naming pattern is utilized to achieve the required operation. Only three Python programs and simple library, used to unify and simplify the implementation of activeMQ and STOMP protocol, are needed to use the framework. A watchdog program is used to monitor, remove, add, start and stop any machine and/or its different tasks when necessary. For every machine a dedicated one and only one zoo keeper program is used to start different functions or tasks, stompShell program, needed for executing the user required workflow. The stompShell instances are used to execute any workflow jobs based on received message. A well-defined, simple and flexible message structure, based on JavaScript Object Notation (JSON), is used to build any complex workflow systems. Also, JSON format is used in configuration, communication between machines and programs. The framework is platform independent. Although, the framework is built using Python the actual workflow programs or jobs can be implemented by any programming language. The generic framework can be used in small national data centres for processing seismological and radionuclide data received from the International Data Centre (IDC) of the Preparatory Commission for the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO). Also, it is possible to extend the use of the framework in monitoring the IDC pipeline. The detailed design, implementation,conclusion and future work of the proposed framework will be presented.