GFZ
Helmholtz Centre
POTSDAM

Rabotnik - An Asynchronous and Flexible Data and Task Management Ecosystem

Felix Delattre, Marius Kriegerowski
and Danijel Schorlemmer

fd@gfz-potsdam.de

HELMHOLTZ CENTRE POTSDAM
GFZ GERMAN RESEARCH CENTRE
FOR GEOSCIENCES

Challenges of Dynamically Changing Big Data

Within the Global Dynamic Exposure (GDE) group at GFZ we enhance earthquake exposure models to a higher resolution by combined processing of different data sources such as the specially pre-processed OpenStreetMap, satellite imagery, cadastral information as well as our own aggregated exposure models for the generation of a higher resolution final exposure model to be used for the assessment of risk due to natural hazards.

To tackle the challenges of Big Data computing in an online operating production system with continuously changing heterogeneous data sources, we designed Rabotnik. It is a Python-based framework for distributed, real-time processing of large datasets, which makes it easy for scientists to define their own processing rules.

Each updated, added or removed datum needs to be reflected within the different analyses which all together constitute the Global Dynamic Exposure models and datasets. The highly complex dependencies between individually developed and maintained data-analysis stages requires a flexible and extensible task queueing, processing and messaging system that is scalable across computational nodes to enable a real-time processing of the ever changing datasets.

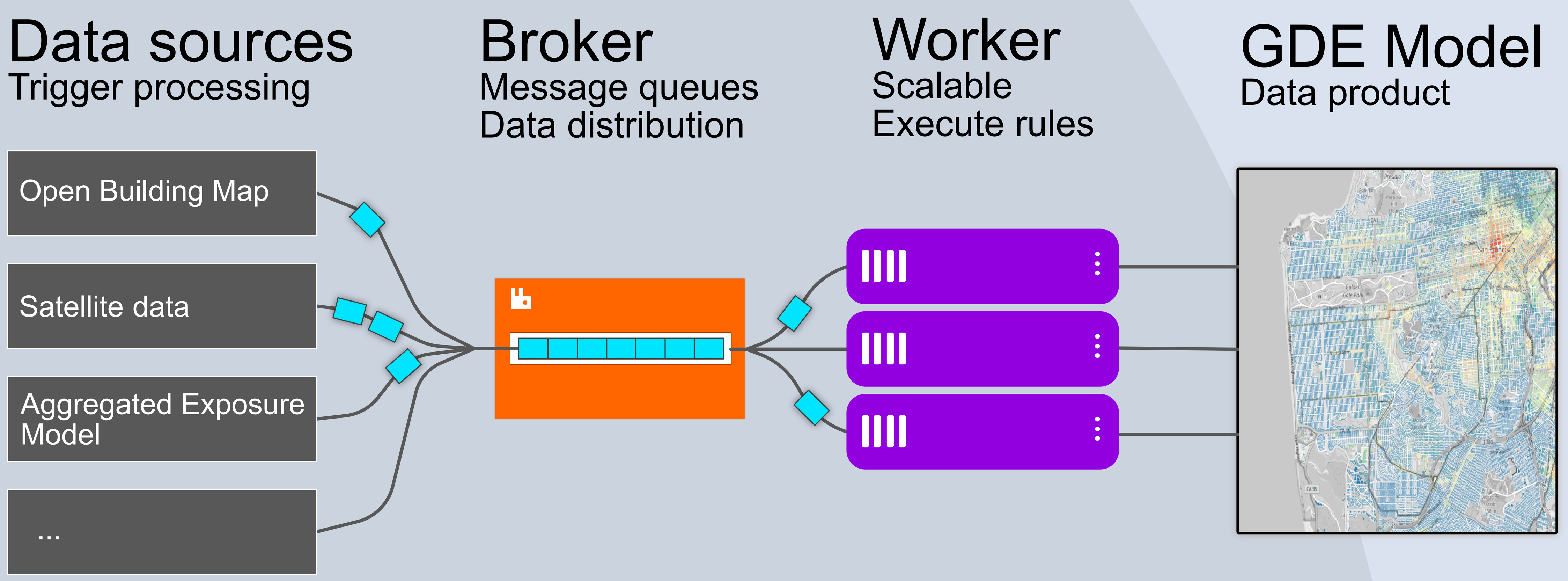
Rabotnik's Main Features

- Job queues and tasks handling: concurrency, priorities, inspect queues
- Scalable distributed processing: Messaging queue between components
- Atomic abstract rules framework: logging, versioning and easy for people to implement. JSON serialization for UI assembling (work in progress).

Foundation

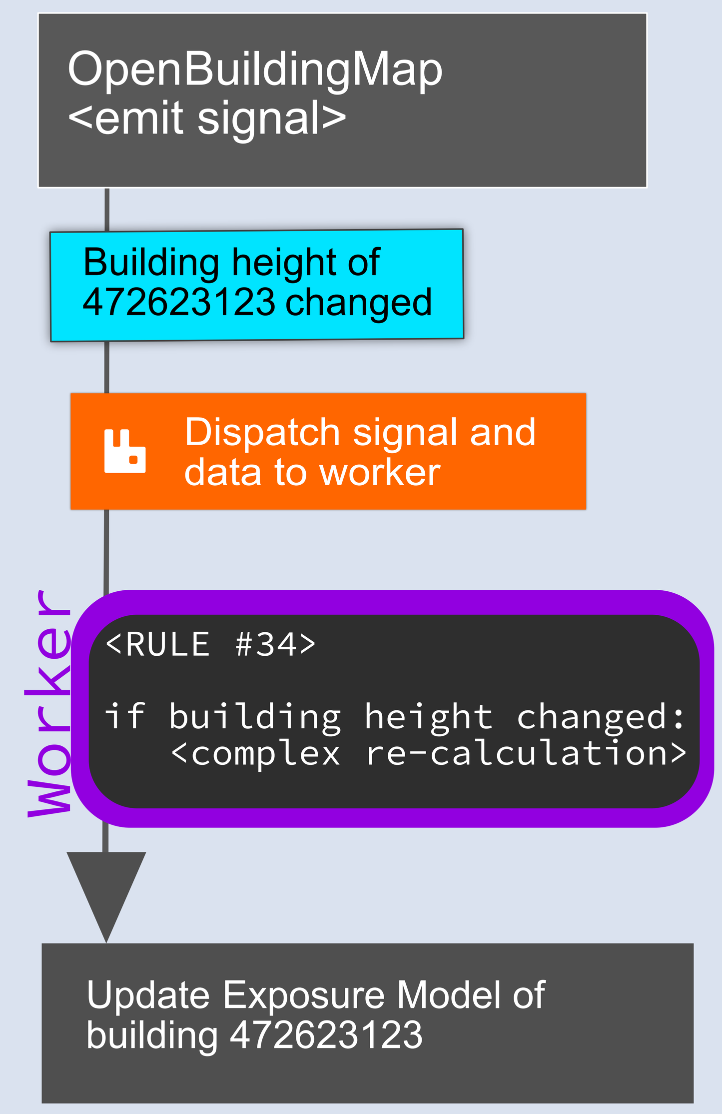
Rabotnik builds on top of existing, solid and state-of-the-art open-source solutions:

- Celery provides distributed processing, within one machine, or distributed over several nodes.
- A message queue for communication between instances conducting processing steps is built with RabbitMQ as a message broker and based on the Message Queuing Telemetry Transport protocol (MQTT) to pass information, triggers and messages across all players that are hooked into a Rabotnik ecosystem.



Rabotnik schematic: Dynamically changing data sources push messages to the broker (rabbitMQ) which enqueues the sent data packages before the worker nodes pull and digest them. The rules executed on the workers encapsulate the logic and feed their results into the Global Dynamic Exposure model.

Data Flow Example



Workflow

The decentralized architecture allows any participant to subscribe to broadcasting channels and to trigger computations, e.g., if a certain datum such as an attribute of a building within the OpenStreetMap database is updated. The receiving instance can in turn issue a message for other subscribed entities to react upon.

Computations can be abstracted in atomic rules which can be dynamically added to the analysis flows. This level of abstraction makes it easier for scientists to program a small logic and add it to the computational system; furthermore it will in the future allow to store rules with a time history of changes which can be recorded in the analysis results to provide transparent and reproducible analyses.

The system is based on asynchronous interaction and processing to leverage a non-blocking code execution flow which is currently being benchmarked by means of several computation rules acting on the OpenStreetMap data to provide a realistic testing scenario with a focus on the European continent.

Code

The current work in progress is published as Free Software, under AGPLv3+ license. And the source code is available at:

https://git.gfz-potsdam.de/dynamicexposure/rabotnik/rabotnik

Contributions, feedback and ideas are very welcome.

Acknowledgement

This work is partially funded by the Real-time Earthquake Risk Reduction for a Resilient Europe (RISE) project (EU Horizon 2020 grant agreement No 821115), the Large-scale EXecution for Industry and Society (LEXIS) project (EU Horizon 2020 grant agreement No 825532), and the Airborne Observation of Critical Infrastructures (Luftgestützte Observation Kritischer Infrastrukturen, LOKI in German) project (German Federal Ministry for Education and Research funding code FKZ 03G0890D).