

Philipp Koster

Graduate Candidate	Philipp Koster
Examiner	Prof. Stefan F. Keller
Co-Examiner	Claude Eisenhut, Eisenhut Informatik AG, Burgdorf
Subject Area	Software and Systems

# Big Spatial Data Analysis and Processing



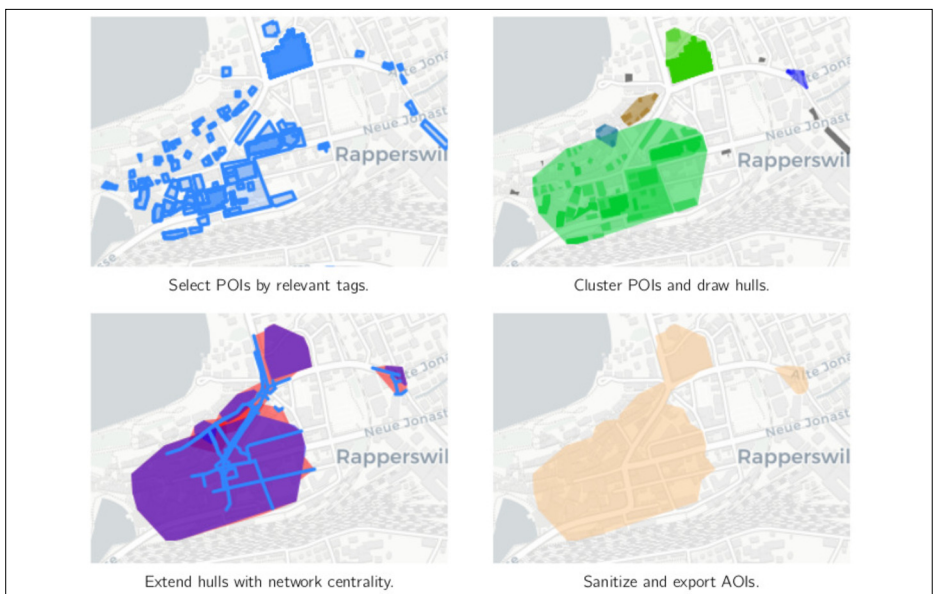
Technology stack of both implementations (first above / second below)

**Introduction:** More and more spatial data is being collected in various domains. To process massive amounts of data, new big data technologies have evolved. These technologies partition and distribute the calculations to the worker nodes of a cluster. The present thesis deals with the evaluation of PostgreSQL/PostGIS and Apache Spark in the processing of massive amount of spatial data while keeping the advantages of SQL in mind. With this in view, two use cases are implemented in the thesis. The first use case involves generation of Areas-of-Interest (AOIs) which are being defined here as "Urban area at city or neighbourhood level with a high concentration of Points-of-Interests (POI) and typically located along a street of high spatial importance". AOI help to explore the world as a traveller. The second use case is about aggregating events consisting of point coordinates, location accuracy and timestamp within polygons. This serves e.g. mobile geotargeting.

**Procedure / Result:** The first use case, AOI, has been implemented using the database PostGIS. A Jupyter notebook was used to elaborate the following process: First, POIs from OpenStreetMap are clustered. Then, geometric hulls of the clusters are extended using a network centrality algorithm. The generated AOIs have been published as a web application and as a GeoJSON file. The second use case, location aggregation, was defined by an industrial partner who reported performance issues with PostGIS. In a first approach, their implementation was optimized and the performance could be improved to their satisfaction by applying diligently sequenced spatial SQL queries. To process geospatial data with Apache Spark, the extension GeoSpark was evaluated over alternatives like GeoMesa. While implementing both use cases, various pitfalls and bugs were encountered in GeoSpark. Additionally, GeoSpark does not yet contain the functionality needed to implement AOIs. Therefore only a small subset of the PostGIS implementation could be implemented with GeoSpark.



Resulting Areas-of-Interest for Zürich (polygons in shaded orange)



The Areas-of-Interest generation process in four steps from upper left to lower right