Hibernate Search

JSR 352 batch job for re-indexing entities

Mentor : *Gunnar Morling* | Student : *Mincong Huang*

Abstract

Hibernate Search is an extension to Hibernate ORM that brings the powers of full-text search via Apache Lucene and Elasticsearch to JPA models. Think of it as Google: after processing the input keywords (e.g. ignoring stop words and applying different kinds of normalization), Hibernate Search returns results ordered by relevance. It brings significant benefits comparing to normal SQL for search-related uses cases. Its full-text index can be updated in two ways: automatically, whenever data changes (using listeners hooked into Hibernate) or by means of the "mass indexer", which rebuilds the entire index for given entity types.

This proposal aims to provide an alternative to the current mass indexer implementation, using the Java Batch architecture as defined by *JSR 352.* This standardized tool JSR 352 provides task-and-chunk oriented processing, parallel execution and many other optimization features. This batch job should accept the entity type(s) to re-index as an input, load the relevant entities from the database and rebuild the full-text index from these.

Proposal Timeline

- **Understanding context** (April 01 - May 15)

  Learn about Hibernate Search, Apache Lucene, Elasticsearch, JBeret, WildFly and other technologies which are relevant to this proposal. Exchange with Hibernate team to ensure a well-understanding of context. Do some simple technology spikes to become familiar with different technologies. Get knowledge about concepts such as indexation, the Query DSL, analyzers, clustering etc.

- **JSR 352 implementation and tests** (May 16 - June 30)

  Implement a JSR 352 batch job for re-indexing entities and run integration tests in plain Java SE (using JBeret as the batch runtime) and JBoss WildFly. Compare the performance of the batch job with the existing mass indexer and find the right tuning parameters to ensure a great performance. Write mid-term evaluation.

- **Enable selection of entities to be re-indexed through different methods** (July 01 - July 31)

  Allow to select the entities to be re-indexed through Criteria or HQL/JPQL query (allowing e.g. to re-index only entities modified after a specific date). Allow to specify a JPA 2.1 entity graph for loading the entities to be re-indexed.

- **Finish features and documentation** (Aug. 01 - Aug. 31)

  Finish merging features. Further refine tests and documentation for the whole project.