Secondary Index

Table of Contents:

Search Engine

Search engines are NoSQL database management systems dedicated to the search for data content. In addition to general optimization for this type of application, the specialization consists in typically offering the following features:

- Support for complex search expressions
- Full text search
- Stemming (reducing inflected words to their stem)
- Ranking and grouping of search results
- Geospatial search
- Distributed search for high scalability

Search Engines are used in OpenCGA as a complementary engine for improving the performance of some queries and aggregations, full *text search* and *faceted* queries to Variant database.

Apache Solr

Apache Solr 6.x is highly reliable, scalable and fault tolerant NoSQL database, it provides distributed indexing, replication, load-balanced querying, automated fail over, recovery, centralised configuration and more.

Currently, the only implementation at OpenCGA uses Apache Solr as Search Engine.

Elasticsearch

Elasticsearch is a distributed, RESTful search and analytics engine capable of solving a growing number of use cases. As the heart of the Elastic Stack, it centrally stores your data so you can discover the expected and uncover the unexpected.

Index with Search Engine

```
opencga-analysis.sh variants secondary-index --project <project>
```

Variants scheme

The goal is to improve the performance of complex queries helping the current storage engine, not to replace the storage engine. There is no point on loading the whole database in the search engine and duplicate all the data. Only a subset of fields is stored, a summary of the annotation and variants structure. This keeps controlled the size of the database, and maintains a manageable dataset.

Most of the Variant queries use filters over VariantAnnotation.

Query intersection

https://github.com/opencb/opencga/issues/638

Faced queries

https://github.com/opencb/opencga/issues/556

Approximated count

https://github.com/opencb/opencga/issues/638

https://github.com/opencb/opencga/issues/749