

# Similarity Join

Filip Peterek

VSB - Technical University of Ostrava

March 2022

Results

Iterations

Current State

# Results



Figure: Best result achieved by solution

# All-Pairs

- ▶ All-Pairs algorithm
- ▶ Inverted Index
  - ▶ Separate index for each record size
- ▶ `robin_hood::unordered_flat_map`

# A Change of Utmost Importance



**Removed all mentions of MS Windows from README**

Filip Peterek authored 1 month ago

**Figure:** Most important improvement

# Parallelization

- ▶ Two different approaches
  - ▶ Multiple queries in parallel
  - ▶ Parallel passes through multiple inverted indices
  - ▶ The latter proved to perform better
- ▶ Parallel data loading
  - ▶ Open source blocking queue

# Lower Level Optimizations

- ▶ Fewer random accesses to memory
  - ▶ One index for a range of sizes
  - ▶ Index range depends on threshold
- ▶ Fewer allocations
- ▶ Iterators over `operator[]`

# Index Size Ranges

---

```
1  /* Higher constant - smaller splits */
2  constexpr double splitConst = 0.7;
3  const double splitRatio =
4      threshold + (1 - threshold)*splitConst;
5  const auto begin = last+1;
6  const auto end = std::min(
7      std::uint32_t(begin / splitRatio),
8      50'000u
9  );
```

---



# Current State

- ▶ Index created before applying All-Pairs
  - ▶ Performance has been decreased slightly