Abstract

One of the widely postulated theory is â€œMost of the time spent working with real world data is not spent on the analysis, but in preparing the dataâ€, I believe any Data guy would agree. There are numerous problems, which crop up while cleansing any dataset, and the prominent and recurring problem is duplicates (It is duplicate, so it has to be recurring). The problem of matching data and information from multiple databases or sources is also a prominent problem encountered in large decision support applications in large commercial and government organizations. This problems has many many different names deduplication, record linkage, entity resolution, coreference, reference reconciliation, record alignment.

Accurate, Scalable and Fast entity resolution has huge practical implications in a wide variety of commercial, scientific and civic domains. Despite the long history of work on Data Matching, there is still a surprising diversity of approaches, and lack of guiding theory. Meanwhile, in the age of big data, the need for high quality entity resolution is growing, as we are inundated with more and more data, all of which needs to be integrated, aligned and matched, before further utility can be extracted.

This talk will present the key ideas implemented behind â€œDedupeâ€ an open source python library that quickly deduplicates and matches records at the scale of millions of records on the laptop. The aim is to show how â€œDedupeâ€ achieves speed by â€œBlockingâ€ records, to save from O(n2) comparisons, achieves accuracy, by using better string comparators and clustering algorithms suited for this problem etc. The attendees would also gain understanding of the tradeoffs between the speed and accuracy.

But what about a billion records ? In such a scenario, it is imminent to parallelise the whole process, to achieve greater speed. So, enter MapReduce based Entity Resolution. Attendees would also walk away with the understanding of how the Deduplication procedure may be parallelised by distributing the task independently to the map and reduce stages. There would also be a demo of the same using â€œDedoopâ€ an open source Efficient Deduplication tool for Hadoop on Amazon EC2 Machines

Outline

The Presentation is multifaceted, its outline/timeline would be:

- Introduction
- Current Real World Data Problems
- Why the Problem is hard ?

(5 mins)

- Current Industry Workflow
- What is Deduplication and Data Matching ?
- â€œDedupeâ€ - Scalable Library for Data Matching in Python

(5 mins)

- Scale Data Matching to millions of Records on your Laptop - DEMO

(5 mins)

- What about Billions of Records ?
- MapReduce based Entity Resolution
- â€œDedoopâ€ - Efficient Deduplication with Hadoop

(5 mins)

- Demo of Deduplication of on a products dataset on Amazon EC2 Machine using dedoop

(5 mins)

- The Road Ahead
- Any questions ?

(5 mins)