Low Latency Geo-distributed Data Analytics – Public Review

Mohammad Alizadeh
MIT Computer Science and Artificial Intelligence Laboratory
alizadeh@csail.mit.edu

Large cloud service providers ingest massive amounts of data in geographically distributed sites spread across the globe. Analytics for such *planetary-scale* datasets is an important emerging challenge. The current practice is to copy all data to a central location, where it can be dealt with locally by standard data analytics stacks such as Hadoop and Spark. However, transferring large volumes of data over the WAN is very costly and slow, and may not even be possible in certain cases due to data sovereignty concerns and legal restrictions.

In the present paper, Pu et al. develop Iridium, a system for low latency analytics across geographically distributed datasets. Iridium explicitly considers WAN bandwidth limitations and jointly optimizes task and data placement across sites to minimize query response times. At a high level, the idea is to move data out of bottleneck sites (with slow WAN links) before queries arrive as well as place tasks during query execution to avoid slow data shuffles across the WAN.

Optimal joint data and task placement is intractable, so the paper instead develops an efficient greedy heuristic. At the heart of the algorithm is a clever use of a simple and efficient LP formulation for optimal task placement given a fixed data placement. The system repeatedly solves the LP to explore many "what-if" scenarios quickly and identify good candidates for moving data out of the bottleneck sites.

The core technical problem Iridium solves has its heritage in distributed databases research dating back more than 30 years. However, the combination of diverse workloads (e.g., Iridium supports arbitrary DAGs of computation), scale, and WAN idiosyncrasies make for interesting new challenges. Indeed, a primary contribution of the paper is in developing simple practical techniques to deal with conflicting requirements and uncertainties in WAN environments. For example, observing that minimizing WAN usage (which some recent work has focused on) is not always well-aligned with minimizing query latency, the authors introduce a "bandwidth budget" in Iridium to balance between WAN costs and query performance.

Iridium was evaluated with a full system implementation on top of Spark, which the authors have made publicly available (`https://github.com/Microsoft-MNR/GDA`). Experiments with several production benchmarks show significant gains (e.g., 3–19× reduction in query latency) relative to a centralized baseline and unmodified Spark.

Iridium's design does have its limitations. The greedy heuristic for task and data placement can get stuck in local optima; thus, it is possible that more sophisticated search algorithms could find significantly better solutions in some cases. Also, while Iridium's particular combination of techniques deliver strong gains in the evaluated workloads, it is unclear how they fare in general. For instance, Iridium prioritizes data movements for queries that arrive sooner but cannot guarantee the move completing before the query arrives. Could this backfire with small query lags or tight bandwidth constraints? Another example is the bandwidth budget: Iridium may end up using the entire budget even if the improvement in query latency is negligible. An alternative approach might make the tradeoff explicit in the algorithm by introducing a "value-of-time" parameter.

Taking a step back, besides better heuristics, an important next step would be to develop performance bounds to guide the exploration and evaluation of different heuristics. Another interesting avenue is to incorporate richer models in the formulation. For example, a more realistic model of WAN pricing could include different cost profiles for different links to capture the complexities of peering and transit arrangements with ISPs.

In summary, geo-distributed analytics is an important emerging area. This paper scopes out part of the design space and develops a practical system backed by a solid implementation and impressive experimental results. Hopefully, the paper will blaze a trail for future research on geo-distributed analytics systems.