**Review of BlinkDB**

# **Summary**

BlinkDB is a distributed sampling-based approximate query processing system that manages to achieve a better balance between efficiency and generality for analytics workloads. It allows users to pose SQL-based aggregation queries over stored data, along with response time or error bound constraints.

**Contributions**

① Using a column-set based optimization framework to compute a set of stratified samples.

② Creating *error-latency profiles* for each query at runtime to estimate error or response time on each available sample.

③ Integrating approach into an existing parallel query processing framework(Hive) with minimal changes.

**Comments**

BlinkDB is based on two important ideas: ① a multi-dimensional sampling strategy that builds and maintains a variety of samples, ② a run-time dynamic sample selection strategy that uses parts of a sample to estimate query selectivity and chooses the best samples for satisfying query constraints.

However, a model that assumes the wrong kind of similarity will lead to a system that “over-fits” to past queries and produces samples that are inflictive at handling future workloads.