# Interactive SQL Queries on Streaming Data (AMS 560 : Big Data Systems, Algorithms and Networks)

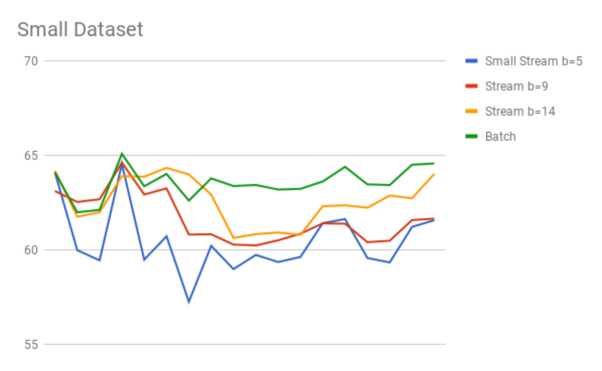
## Problem

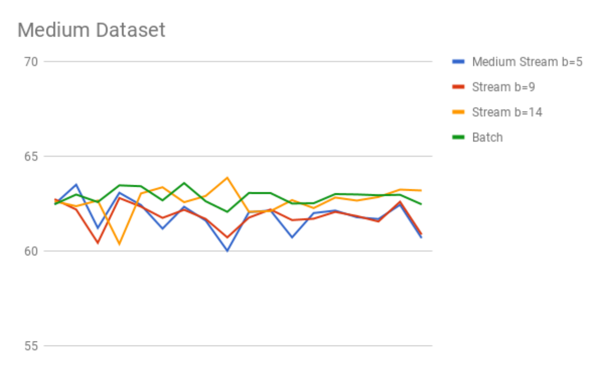
- Answering SQL queries on Streaming data using Apache Spark in real time
- Naive method to execute this is of quadratic complexity

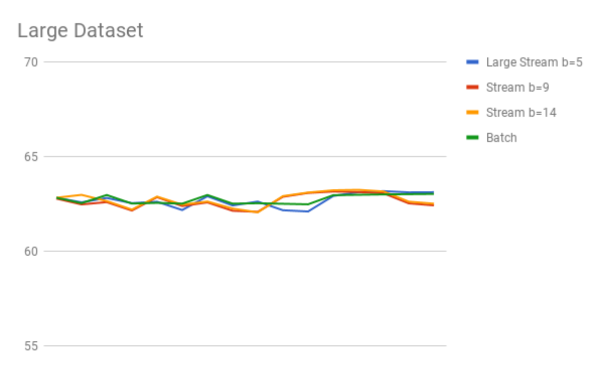
## Key Idea

- Implemented a solution based on idea of G-OLA (a mini-OLA execution model) in PySpark to answer the query in real time
- Each RDD of DStream is considered as a partition and the intermediate output is used to determine the uncertain and deterministic sets
- An approximate answer is outputted after every user specified time interval
- Using State maintenance techniques and bootstrap, we answer user defined SQL queries in real time
- $Q(P_i) = Q(P_{i-1}) + \Delta Q(P_{i-1}, \Delta P_i)$ where, $P_i$ denotes $i^{th}$ RDD and $Q(P_i)$ denotes the output of SQL query after $i^{th}$ RDD

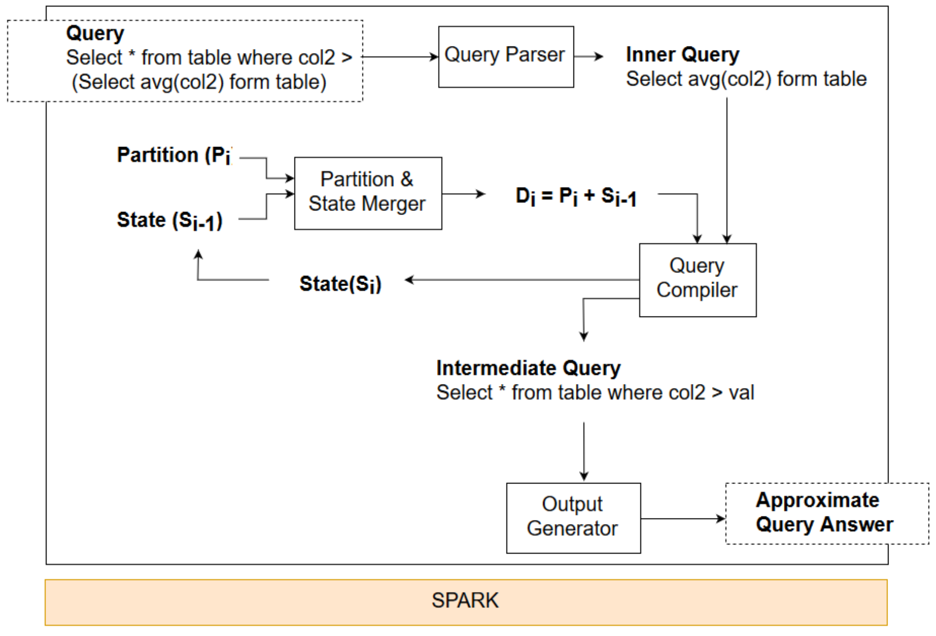
## Results


Small Dataset


Medium Dataset


Large Dataset

## System Architecture



## Code

https://github.com/alok123t/AMS560-Project

## Future Work

- Support for more aggregate operators
- Implementing the bootstrap model for robustness
- Efficient temporal based state maintenance

## References

- Kai Zeng et al. G-OLA: Generalized On-Line Aggregation for Interactive Analysis on Big Data (SIGMOD ’15) https://doi.org/10.1145/2723372.2735381
- Sameer Agarwal et al. BlinkDB: queries with bounded errors and bounded response times on very large data (EuroSys ’13) http://dx.doi.org/10.1145/2465351.2465355
- Matei Zaharia et al. Discretized streams: fault-tolerant streaming computation at scale (SOSP ’13) https://doi.org/10.1145/2517349.2522737
- Matei Zaharia et al. Spark: cluster computing with working sets (HotCloud’10)

**Alok Thatikunta, Sweta Kumari, Rahul Sihag**
{ athatikunta, swkumari, rsihag } @cs.stonybrook.edu

Stony Brook University