ABSTRACT

In the context of data-series, the volume of data in applications such as biology, astronomy, and health care is tremendous. The time spent on explorative queries can be time prohibitive for the user as queries are usually executed to completion and then returned, i.e. all or nothing queries. Another pitfall of most large-data systems and explorative data analysis is a general lack of incremental interaction between the data-system and the data scientist. Providing intuitive visualizations that also allow for the user to not only interact with the already processed data but also influence how the data-system proceeds.

Our project is a prototype system that provides the basic functionality from end to end, data-processing of large, unstructured files, basic statistical processing, a visualization tool capable of handling a stream of summaries, and a feedback mechanism. This prototype explores the trade-offs between large-file streaming and processing with real-time visualization and user interaction with that data. We hope that the prototype will generate excitement for more real-time, interactive data-processing mechanisms, making the exploration of large data less intimidating and tedious and possibly even fun!

MOTIVATION

- The growth of sensors and the ensuing data deluge make it a necessity for tools that scale
- Data exploration over time or data-series is imperative
  - locating regions of interest,
  - changes in modalities (stock market shifts, changes observed in blood pressure and heart rate changes, etc.)
- Exploration can be aided by interactive visualization
- A user’s interaction with the visualization can help inform how the data processing system proceeds

DATA STORAGE AND PROCESSING

- Data in large volumes is typically stored in unstructured formats
- Loading data and building indexes is prohibitive for exploration
- Need low overhead, fast scanning with a visualization component!
- In memory structures built for visualization and handling feedback

VISUAL INTERFACE

- Visualization will allow interaction from the user in terms of showing sampling regions
- Allow the user to indicate how sampling should proceed
- Show summary information of all current columns through histograms
- Interface easy for new functionalities such as clustering, correlation between dimensions, etc.

REFERENCES