Using Naiad to Analyze Twitter Data in Batch and Real-time

George Wort
University of Cambridge
2017
Naiad

• Timely Dataflow System.
  • Batch Processing.
  • Stream Processing.
  • Graph Processing.
• Supports iterative and incremental data analysis.
• Low latency.
• High throughput.
Naiad

- Complex system offering a lot of options.
- Too complex for most applications?
- Overheads and ease of use?

- Additions:
  - Differential Dataflow.
  - GraphLINQ.
Twitter Data Processing

- Implement real-time and batch processing of tweet stream.
- Geographically categorise word frequencies.
- Allow selection of different levels of granularity.
- Query geographical data.
- Extend to allow similarity comparison between areas or cluster areas in batch.
- Extend to view frequency of spelling mistakes in English.
Assessment

• Implement on a single machine and distributed environment.

• Using:
  • The base Naiad system.
  • Differential dataflow.
  • GraphLINQ.

• Assessing:
  • Ease of use.
  • Flexibility.
  • Latency.
  • Throughput.
Questions?