High Performance Streaming Analytics
Break Out Group

Alison Brizius, Cees de Laat, David Ediger, Melvin Ramos, Mike Papka, Nick Holliman, Robert Grossman

AoT workshop, Chicago, September 4, 2015
Streaming Analytics

- A data stream is an ordered sequence of data records that can be read only once, with only limited computing and storage capabilities available for processing.
- Think of this as “infinite” data, finite storage, read once.
- Compare to a batch analytics system or HPC system in which data can be read as often as required to complete the analysis.
- There are some research and several commercial systems that can process data streams.
- Over the past several years, there have been some recent open source frameworks (Akka, Storm, etc.) for stream processing that have been used for production deployments.
- There is relatively little work done on streaming analytics systems and even less on high performance streaming analytics systems.
Distributed sensors with in situ processing

Access and analysis by systems and sensor developers

“Early access” by app developers, researchers, public, etc.

Access to QC and archived data from a repository

Current model under discussion
High performance streaming analytics systems

- Distributed sensors with in situ processing
- Access and analysis by systems and sensor developers
- “Early access” by app developers, researchers, public, etc.
- Access to QC and archived data from a repository

Streaming analytics system

- Derived data products in real time
- Outputs of streaming analytic models
- Alerts, etc.
Streaming Analytics Use Cases

- Assume 10,000+ sensors distributed in a city
- What we can do with today’s technology
  - Continuously updated heat map
  - Continuously pollution/air quality map
  - Continuously updated wind chill, noise pollution, etc.
  - Real time anomaly detection
- What we would like to do, but cannot do today
  - Complex data fusion in real time
  - Integrating the results of ad-hoc simulation into data products derived from streaming analytics
  - Streaming “re-analysis" of complex data streams
  - Smart trade off between in situ processing at sensors, centralized streaming processing and batch processing
    - Variant is regional streaming processing and roll up
High performance streaming analytics model

Distributed sensors with in situ processing → Access and analysis by systems and sensor developers → “Early access” by app developers, researchers, public, etc. → Access to QC and archived data from a repository

- Could we design a specialized HP streaming analytics system?

- Derived data products in real time
- Outputs of streaming analytic models
- Alerts, etc.

- Many of us have designed high performance computing systems.
- Would it make sense to consider how to design a specialized system for high performance streaming analytics that could be used to produce real time streaming analytics over a smart city?
UvA- GPS Bird tracking

- Solar powered GPS logger 12 – 18 g
- 50 - 10,000 measurements per day
- Bi-directional wireless data transmission
- Upload new measurement schemes
- Base station & (mobile) relay network
Shamoun-Baranes et al 2011