Scalable Data Processing Framework

Motivation
- Current Approach for Using Data
  - Each application needs to implement its own data cleansing (Process Redundancy)
  - Multiple accesses to a shared resource (Data Access Redundancy)

System Architecture
- Built on Microsoft StreamInsight.

Introduction
- Motivation: The quality of upstream raw data can be poor to various errors.
  - We propose a Scalable Data Processing Framework (SDPF) that can address a variety of upstream operating data quality issues.
  - Key features include (1) Online/Real-time processing of data, (2) Configurability and (3) Scalability.

Capability
- Configurable:
  - Users can mix and match various types of operators to create a cleansing plan.
  - Each cleansing plan is equivalent to Directed Acyclic Graph with a source node as the data source and destination node as the data sink.

Related Research
- Data Cleansing and Data Compression
- Real-time data streaming/query/integration: database, performance
- Complex Event Processing
- Dynamic Principal Component Analysis (DPCA)

Conclusion and Future Work
- Perform a field test on tens of thousands of data streams.
- Automate the process of plan generation.

IMSC Retreat 2015