Scalable Data Processing Framework

Mohammad Asghari, Luciano Nocera, Seon Ho Kim, Cyrus Shahabi
Integrated Media Systems Center
University of Southern California

Introduction

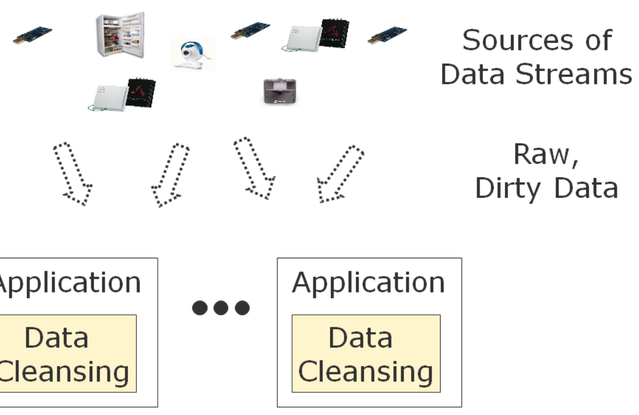
▪ **Motivation:** The quality of upstream raw data can be poor to various errors.

▪ We propose a ***Scalable Data Processing Framework (SDPF)*** that can address a variety of upstream operating data quality issues.

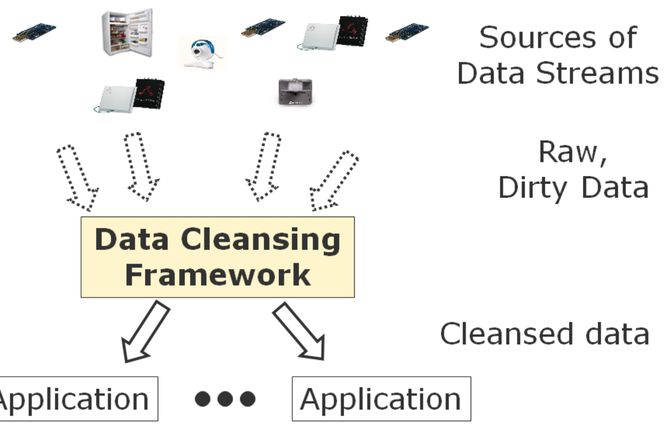
▪ Key features include (1) Online/Real-time processing of data, (2) Configurability and (3) Scalability.

Motivation

▪ Current Approach for Using Data

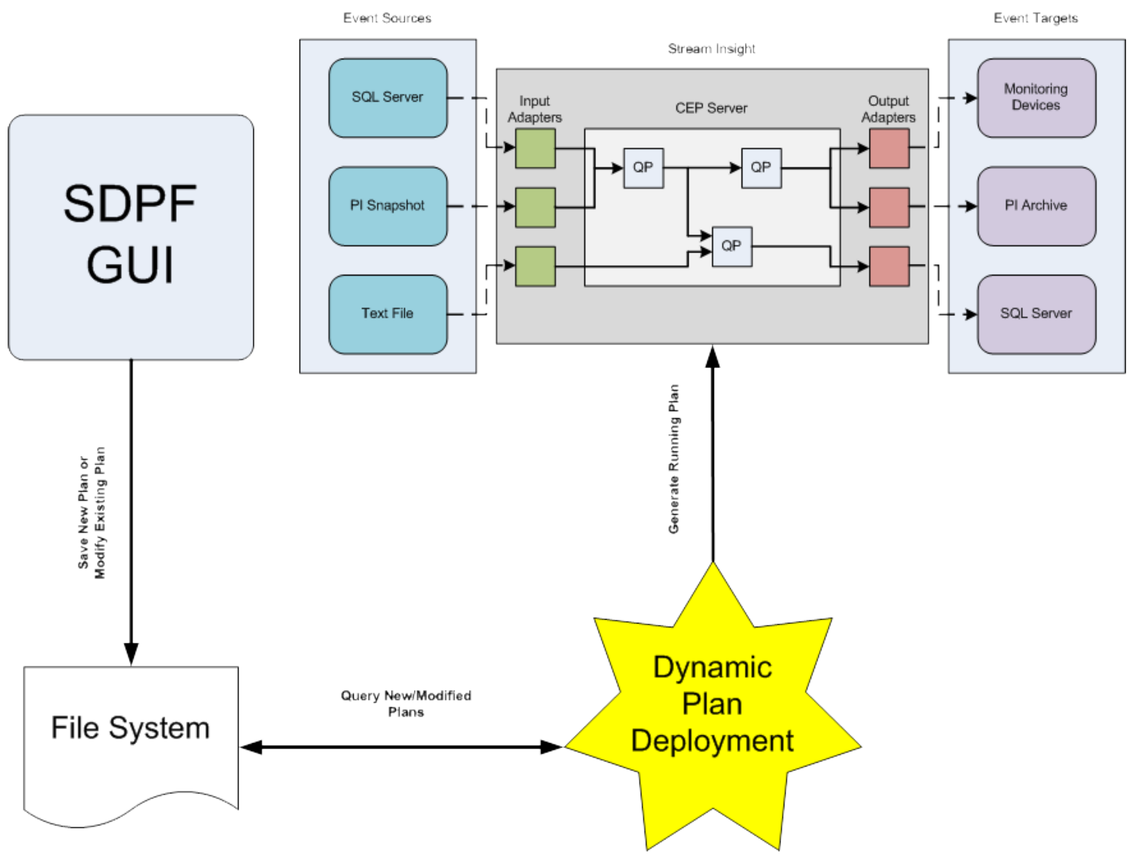


▪ Each application needs to implement its own data cleansing **(Process Redundancy)**
▪ Multiple accesses to a shared resource **(Data Access Redundancy)**



System Architecture

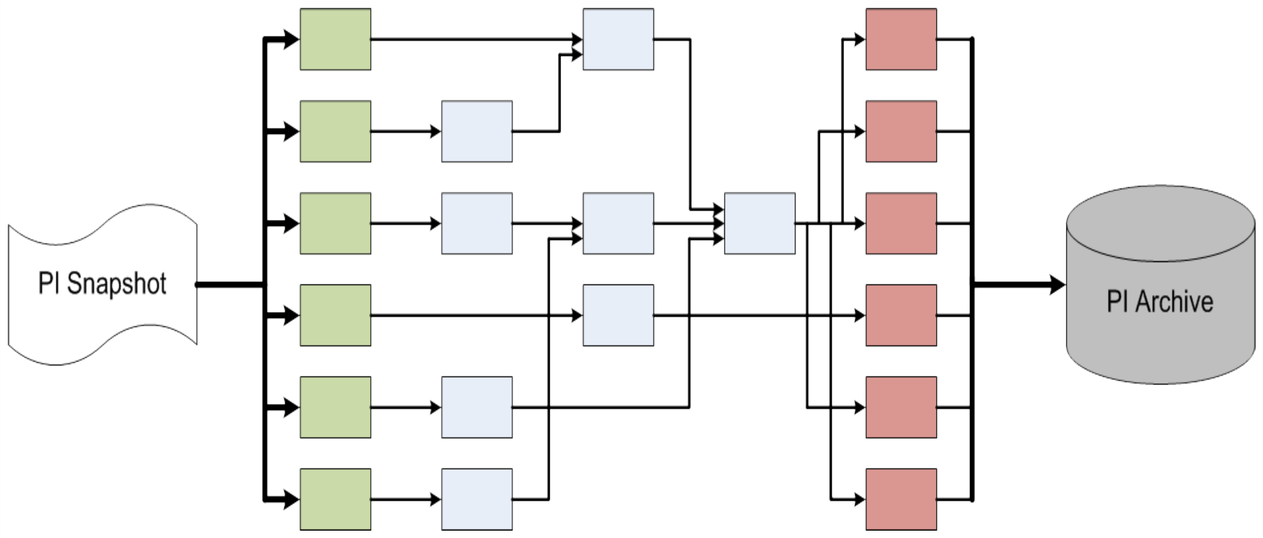
▪ Built on Microsoft StreamInsight.



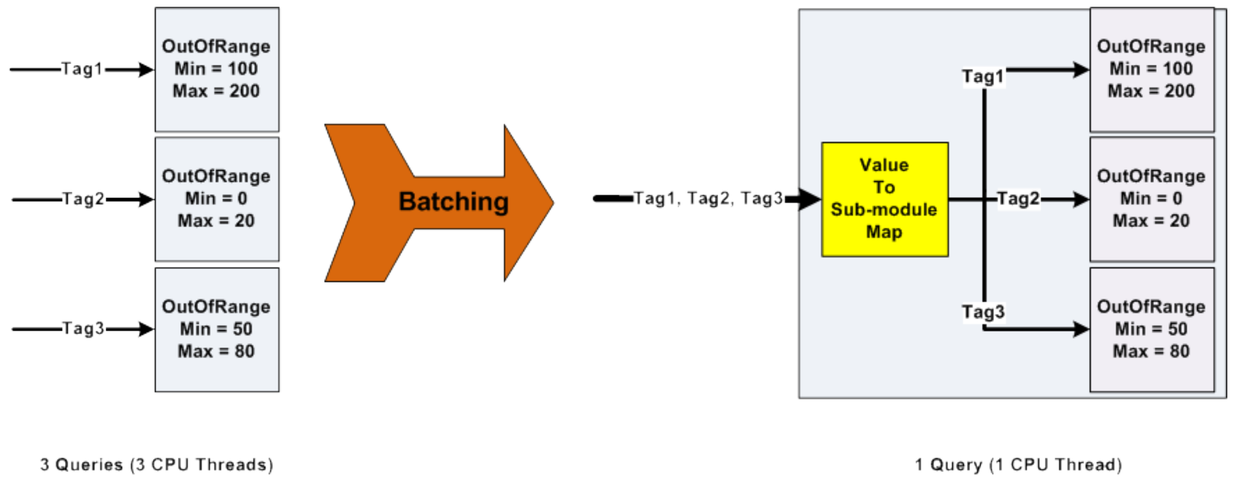
Capability

• **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.



Scalable:
+ By batching similar operators into a single continuous query, the number of systems threads is reduced drastically.



+ The framework also detects different components of the plan that run independently such that each component can run on a single node on the cloud.

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.