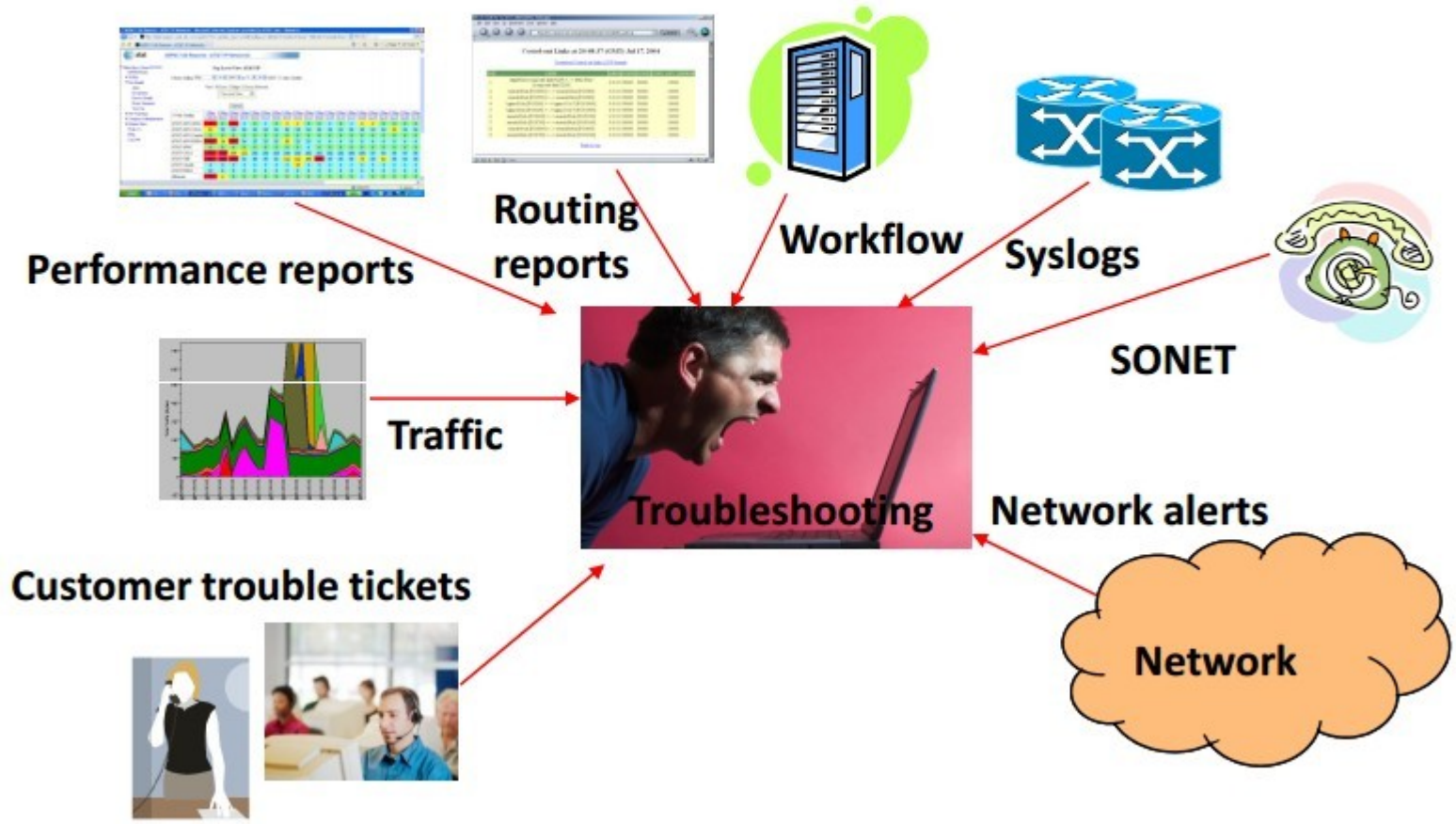
Framework of Data/Event Stream Processing

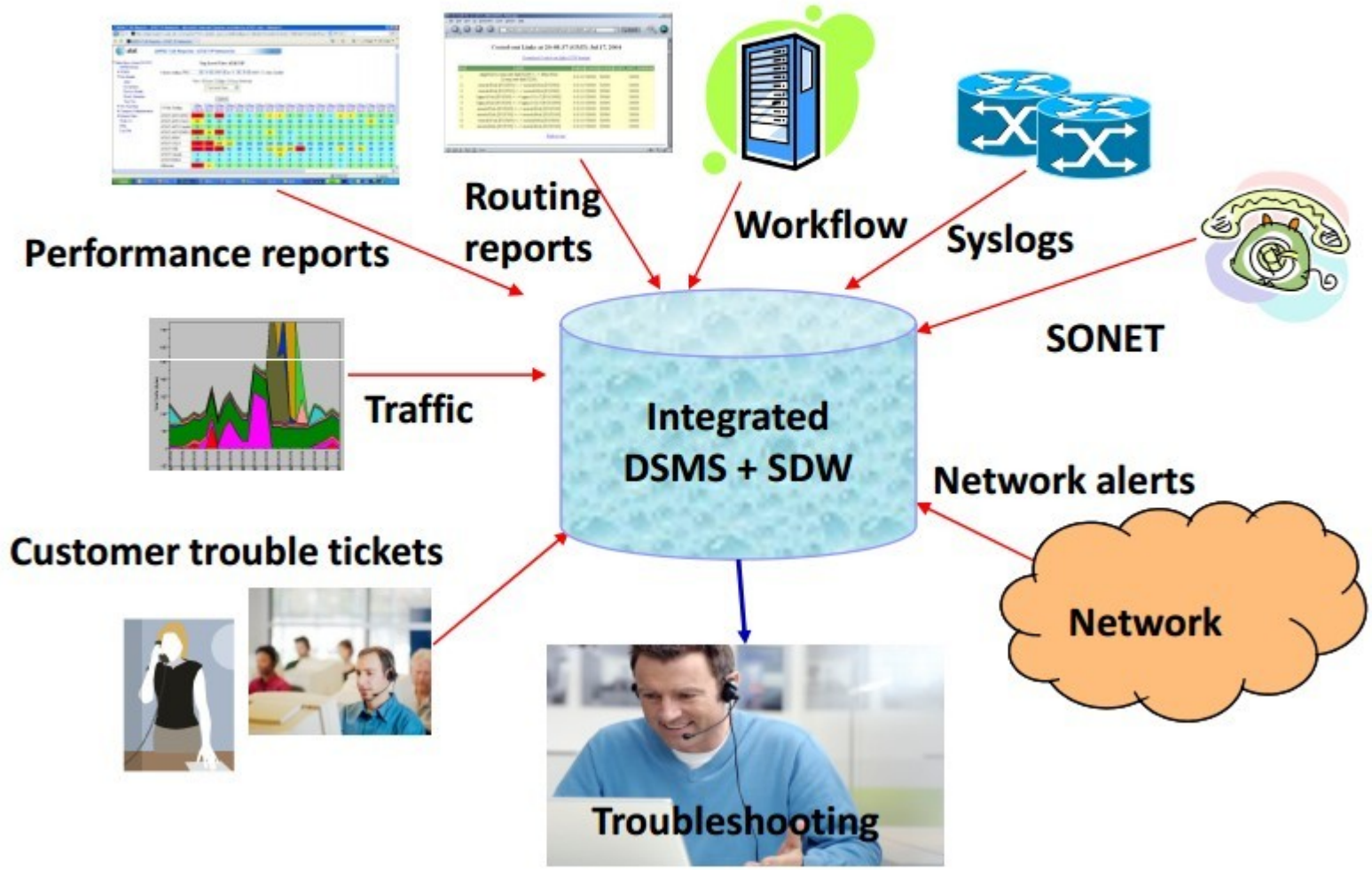
Including monitoring, management, mining, summarization, visualization

Challenge: Enabling Real Time Analysis



- Copy from VLDB 2010 Keynote

Solution: Streaming Data Management

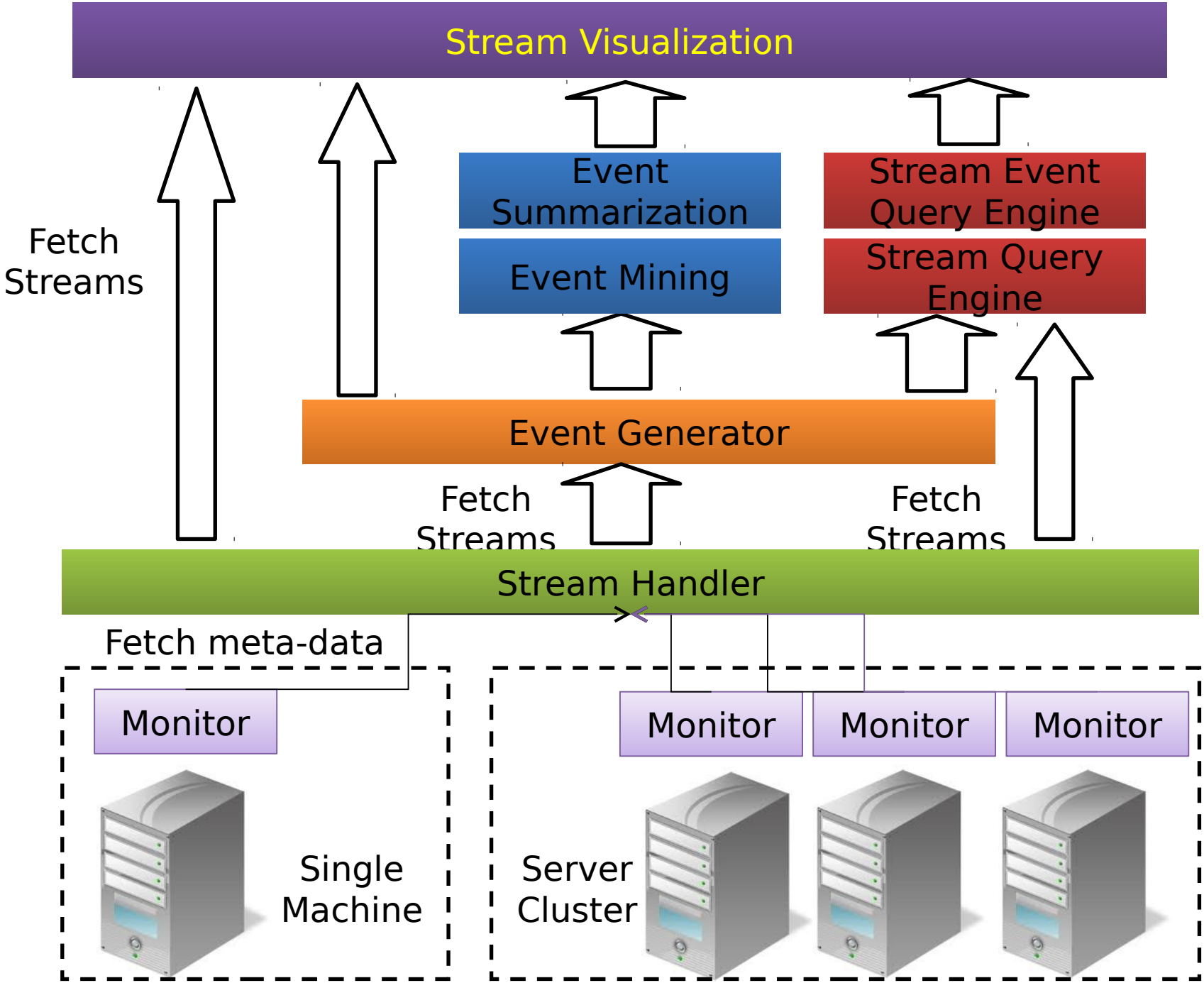


- Copy from VLDB 2010 Keynote

Goal

- Build a complete and flexible framework that all the mining, summarization, query, visualization techniques can be embedded in.

- Interdiscipline research area: data mining, data management, distributed system, service computing, data visualization


Stream Visualization
Event Summarization
Stream Event Query Engine
Fetch Streams
Event Mining
Stream Query Engine
Event Generator
Fetch Streams
Fetch Streams
Stream Handler
Fetch meta-data
Monitor
Monitor
Monitor
Monitor
Single Machine
Server Cluster

(System) Event Monitor

- Automatic monitoring configuration
  - Register monitoring streams
    - CPU, Memory, Network, Cache, Hard Disk
  - Monitor rate configuration
  - Monitor Granularity
    - Total CPU usage, usage per core, usage per process, usage per port

System Monitor

- Challenge: How to monitor as few as possible while capture as much as po ssible useful information.
- Too many monitoring:
  - Increases overhead to system.
  - Brings extra processing burden to followi ng processing modules

Possible Solution

- Estimate the cost of conducting a monitoring action (add/remove monitor stream, change monitor rate)
  - Cost of possible information loss
  - Cost of overhead
- Example:
  - No need to monitor stable streams
  - Need to pay more attention to high resource consumer

Event Generator

- Event should be customized instead of fixed
  - Event generation rule should be set by domain e xpert themselves.
  - Different users has different preferences, ignore uninteresting events
  - Composite Event that generated based on multi ple stream should be support
    - e.g. CPU usage > 95% and memory usage > 95%
    - e.g. hard-disk space of all the servers in cluster is low er than 5%

Event Generator: The Scenario

- Cloud auto-scaling
  - Trigger the auto-scaling mechanism of t he cloud system
- Autonomy Server
  - Shutdown/startup service
  - Clean unused data automatically

Application Level

- Event Mining: Online/Offline
  - Correlation analysis
  - System Diagnosis
- Event Summarization: Online/Offline
- Complex Stream Querying:
  - Continuous querying, Numerical analysis
- Complex Event Querying
  - Continuous querying, relational query
- Stream Visualization
  - Graphics

Reference

- Enabling Real Time Data Analysis. VLDB keynote. 2010.
- Monitoring:
  - Cayuga: A general purpose event monitoring system. CIDR 2007.
- Query & Pattern Matching:
  - TelegraphCQ: An Architectural Status Report. *IEEE Data Engineering Bulletin, Vol 26(1), March 2003*.
  - Efficient Pattern Matching over Event Streams. SIGMOD. 2008.
  - DejaVu: Declarative Pattern Matching over Live and Archived Streams of Events. SIGMOD, 2009.
- Visualization:
  - Visualization and Analysis of Streaming Data. AT&T, JSM 2008.