Survey of Modern Distributed Stream Processing Systems

Milinda Pathirage
School of Informatics and
Computing
Indiana University
Email: mpathira@indiana.edu

Beth Plale
School of Informatics and
Computing
Indiana University
Email: plale@cs.indiana.edu

***Abstract*—Interest on continuous queries over streams of data has increased over last couple of years due to the need of derving actionable information as soon as possible to be competitive in the fast moving world. As a result of the limitations in batch processing technologies from previous generation, distributed stream processing systems like Yahoo's *S4*, Twitter's *Storm*, *Spark Streaming* and LinkedIn's *Samza* were introduced into the fast growing Big Data eco-system. Even though there are various different stream processing platforms and frameworks on top of them with different capabilities and characteristics, in-depth comparative studies of performance, scalability and reliability has never been done. Users of these system often face difficulties when choosing a system as a solution to a task at hand. In this paper we compare and contrast modern distributed stream processing systems – *S4, Storm, Spark Streaming and Samza* – based on "The 8 Requirements of Real-Time Stream Processing".**

I. Introduction

II. Introduction

Introduction to stream processing.

Introduction to distributed stream processing and its importance.

III. Taxonomy of Distributed Stream Processing

A. Algorithms & Applications

B. Scaling

C. Fault Tolerance

D. System Architectures

IV. 8 Requirements of Real-Time Stream Processing

A. Keep the Data Moving

B. Query using SQL on Streams

C. Handle Stream Imperfections (Delayed, Missing and Out-of-Order Data)

D. Generate Predictable Outcomes

E. Integrate Stored and Streaming Data

F. Guarantee Data Safety and Availability

G. Process and Respond Immediately

V. Distributed Stream Processing Survey

References