A Survey of the State-of-the-art in Event Processing

Otávio M. de Carvalho, Eduardo Roloff, Philippe O. A. Navaux

Federal University of Rio Grande do Sul
Parallel and Distributed Processing Group
{omcarvalho,eroloff,navaux}@inf.ufrgs.br

GPPD

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

.inf
INSTITUTO
DE INFORMÁTICA
UFRGS

Motivation

- We currently live in
  a **world** of **2.8 trillion GB**

- **Data production** rate
is **growing 60% by year**

- Every **60 seconds**, **600 new blog posts** are published
and **34,000 tweets** are sent

Applications of Information Flow Processing

Sensor networks, …

traffic engineering, …

social networking, …

financial markets, …

generate streams!

The History Behind Event Processing Tools

- The **lastest advance** in **Event Processing** can be characterized by two main sub-domains: **CEPs** and **DSMSs** systems that were developed in the early **90s**

- **Main problem**: Today's systems present **intersections** between these two sub-domains

- A new **naming convention** was proposed: **Information Flow Processing** (IFP)

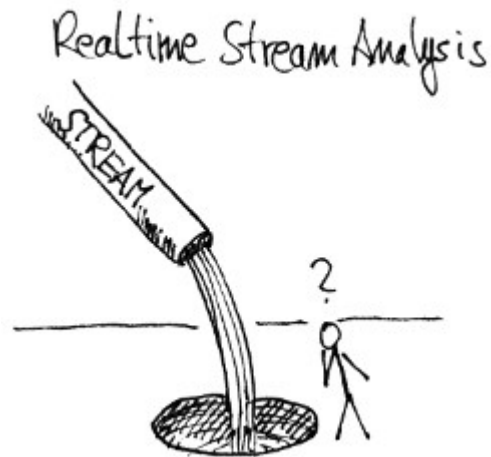| Complex Event Processing Systems	Data Stream Management Systems
- Those systems **focuses** on **processing event** notifications, with a special attention to their ordering relationships, to capture **complex event patterns**.	- Those systems **focuses** mainly on **flowing data** and the application of **transformations** over data

MapReduce is Still Important

- The Google's research on MapReduce model was fundamental to the evolution of event processing systems into distributed event processing systems

- The adoption of Apache Hadoop and his core tools has eased the process of developing distributed systems
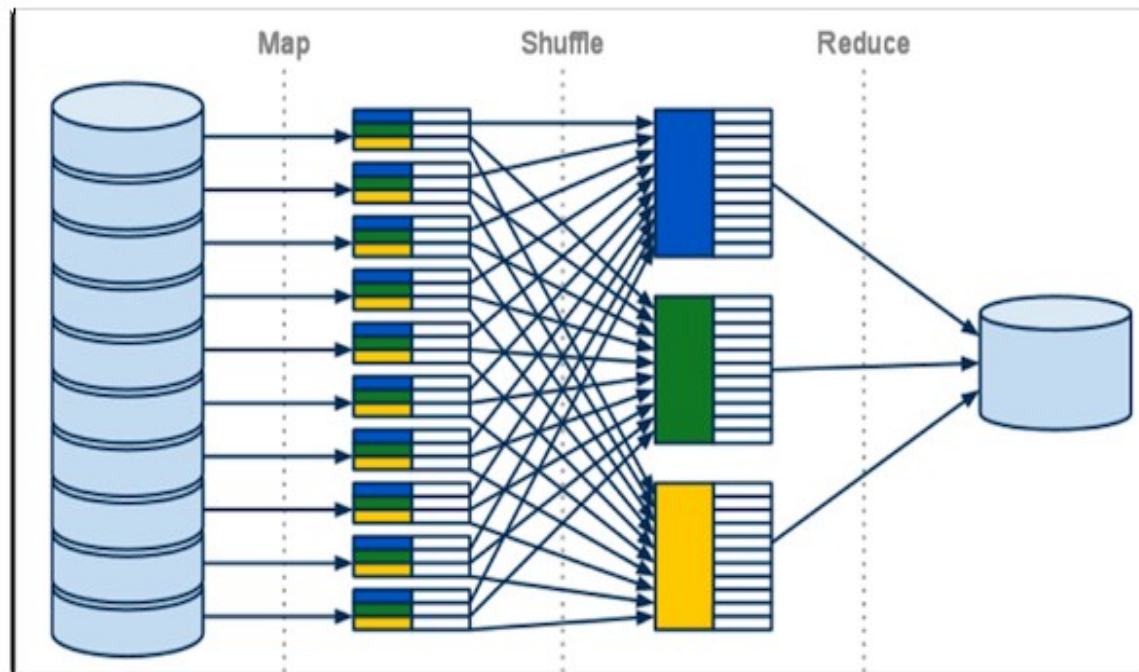
But MapReduce Model Isn't Fast Enought

State-of-the-art
in
Event Processing

Characterization of Tools

Name	Release Year	DBMS	DSMS	CEP	Distributed
Google Photon	2013		•		•
Walmart Muppet	2012		•		•
StreamDrill	2012			•	
SAP HANA	2011	•	•		•
Apache Storm	2011		•		•
Apache YARN	2011	•	•	•	•
Apache Flume	2011		•		•
Apache Kafka	2011		•		•
Apache S4	2011		•		•
Apache Chukwa	2010		•		•
HStreaming	2010		•		•
AMPLab Spark	2010	•	•		•
VoltDB	2010	•	•		•
Esper	2006			•	
StreamBase CEP	2003			•	•
SQLstream	2003		•		•

Characterization of Tools

Name	Release Year	DBMS	DSMS	CEP	Distributed
Google Photon	2013		•		•
Walmart Muppet	2012		•		•
StreamDrill	2012			•	
SAP HANA	2011	•	•		•
Apache Storm	2011		•		•
Apache YARN	2011	•	•	•	•
Apache Flume	2011		•		•
Apache Kafka	2011		•		•
Apache S4	2011		•		•
Apache Chukwa	2010		•		•
HStreaming	2010		•		•
AMPLab Spark	2010	•	•		•
VoltDB	2010		•		•
Esper	**2006**			•	
StreamBase CEP	**2003**			•	•
SQLstream	**2003**		•		•

Legacy Tools

Characterization of Tools

Name	Release Year	DBMS	DSMS	CEP	Distributed
Google Photon	2013		•		•
Walmart Muppet	2012		•		•
StreamDrill	2012			•	
SAP HANA	2011	•	•		•
Apache Storm	2011		•		•
Apache YARN	2011		•	•	•
Apache Flume	2011		•		•
Apache Kafka	2011		•		•
Apache S4	2011		•		•
Apache Chukwa	2010		•		•
HStreaming	2010		•		•
AMPLab Spark	2010	•	•		•
VoltDB	2010	•	•		•
Esper	2006			•	
StreamBase CEP	2003			•	•
SQLstream	2003		•		•

In-memory databases

Characterization of Tools

Name	Release Year	DBMS	DSMS	CEP	Distributed
Google Photon	2013		•		•
Walmart Muppet	2012		•		•
StreamDrill	2012			•	
SAP HANA	2011	•	•		•
Apache Storm	2011		•		•
Apache YARN	2011			•	•
Apache Flume	2011				•
Apache Kafka	2011		•		•
Apache S4	2011		•		•
Apache Chukwa	2010		•		•
HStreaming	2010		•		•
AMPLab Spark	2010	•	•		•
VoltDB	2010	•	•		•
Esper	2006			•	
StreamBase CEP	2003			•	•
SQLstream	2003		•		•

Flow processing

Characterization of Tools

Name	Release Year	DBMS	DSMS	CEP	Distributed
Google Photon	2013		•		•
Walmart Muppet	2012		•		•
StreamDrill	2012			•	
SAP HANA	2011	•	•		•
Apache Storm	2011		•		•
Apache YARN	2011	•	•	•	•
Apache Flume	2011		•		•
Apache Kafka	2011				•
Apache S4	2011				•
Apache Chukwa	2010		•		•
HStreaming	2010		•		•
AMPLab Spark	2010	•	•		•
VoltDB	2010	•	•		•
Esper	2006			•	
StreamBase CEP	2003			•	•
SQLstream	2003		•		•

General-purpose systems

Characterization of Tools

Name	Release Year	DBMS	DSMS	CEP	Distributed
Google Photon	2013		•		•
Walmart Muppet	2012		•		•
StreamDrill	2012			•	
SAP HANA	2011				•
Apache Storm	2011				•
Apache YARN	2011			•	•
Apache Flume	2011		•		•
Apache Kafka	2011		•		•
Apache S4	2011		•		•
Apache Chukwa	2010		•		•
HStreaming	2010		•		•
AMPLab Spark	2010	•	•		•
VoltDB	2010	•	•		•
Esper	2006			•	
StreamBase CEP	2003			•	•
SQLstream	2003		•		•

Specific to solve one problem (top-k problem)

13

Hadoop Influency

Name	Based on Hadoop and his core tools	Interact with Hadoop
Google Photon		
Walmart Muppet		
StreamDrill		
SAP HANA		●
Apache Storm		
Apache YARN	●	
Apache Flume		●
Apache Kafka		●
Apache S4		
Apache Chukwa	●	
HStreaming	●	
AMPLab Spark		●
VoltDB		●
Esper		
StreamBase CEP		●
SQLstream		●

Challenges in Distributed Event Processing

- How to **program**?

- In **what applications** can it be **used**?

- **Now** it **scales**, but is the **throughput** the **same**?

Conclusions

- **MapReduce** paper and later contributions have **collaborated** to the **changes** into **event processing systems**

- **Convergences** between the tools still are causing **naming problems**

- It is not possible to determine if these **implementations** will **converge** into greater sets of tools of general-purpose systems or more specialized systems.

Future Works

- In our future works, we will **select** a **subset of these tools** to **analyze** further

- We will **evaluate** the **performance** of these selected tools

- And finally, we will **apply one** of them in a **real HPC application**

A Survey of the State-of-the-art in Event Processing

Thanks!

Otávio M. de Carvalho, Eduardo Roloff, Philippe O. A. Navaux

Federal University of Rio Grande do Sul
Parallel and Distributed Processing Group
{omcarvalho,eroloff,navaux}@inf.ufrgs.br