Distributed Streaming

Albert Bifet

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

May 2012

COMP423A/COMP523A Data Stream Mining

Outline

1. Introduction
2. Stream Algorithmics
3. Concept drift
4. Evaluation
5. Classification
6. Ensemble Methods
7. Regression
8. Clustering
9. Frequent Pattern Mining
10. **Distributed Streaming**

Data Streams

Big Data & Real Time

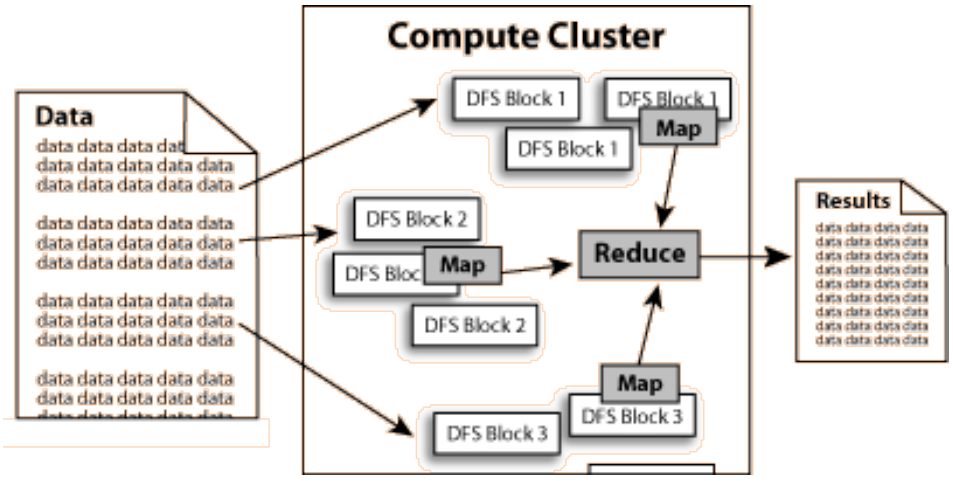
Distributed Systems

Hadoop, S4 and Storm

Hadoop

Hadoop

Hadoop

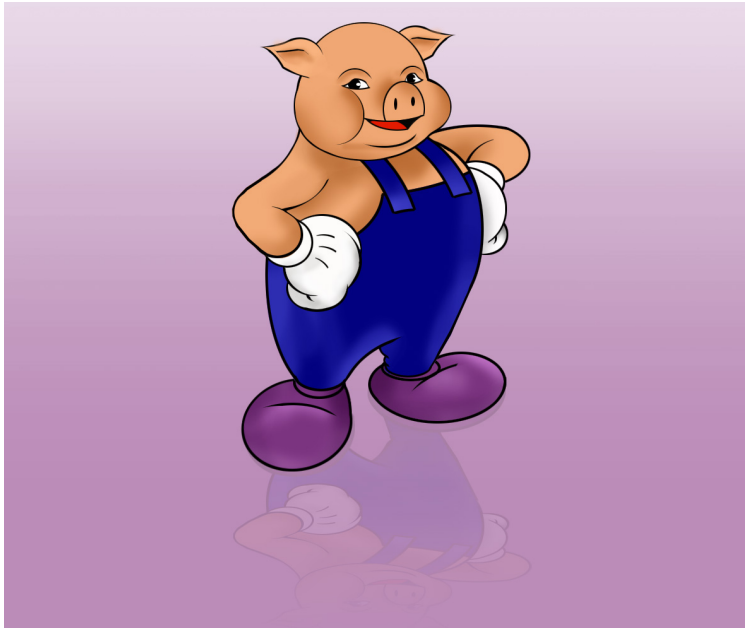


Hadoop architecture

Apache Mahout

Mahout: open source framework

Pig



Pig: Similar to SQL

Pig

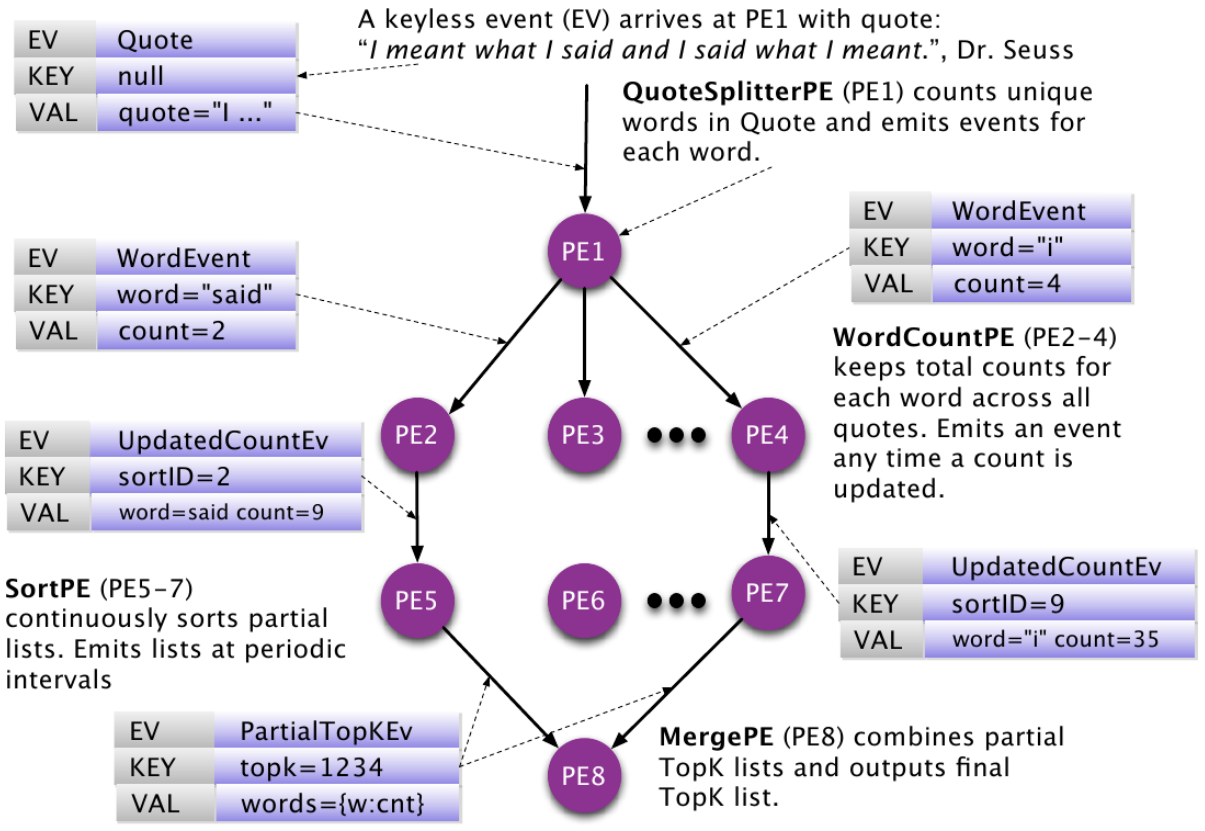
- `A = LOAD 'data' USING PigStorage() AS (f1:int, f2:int, f3:int);`
- `B = GROUP A BY f1;`
- `C = FOREACH B GENERATE COUNT ($0);`
- `DUMP C;`

Pig: Similar to SQL

Apache S4

Apache S4

Apache S4

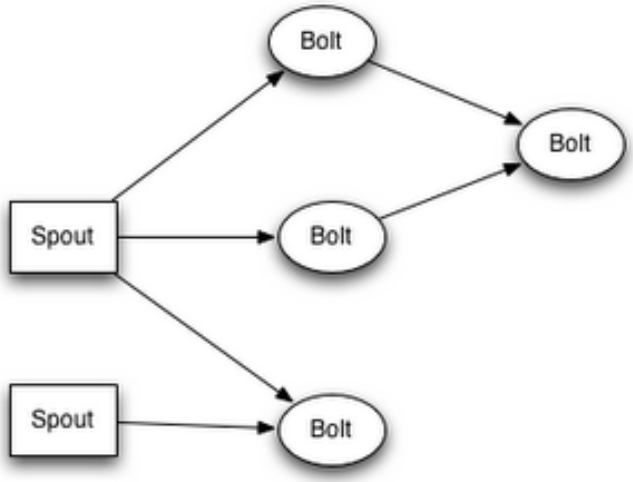


PE ID	PE Name	Key Tuple
PE1	QuoteSplitterPE	null
PE2	WordCountPE	word="said"
PE4	WordCountPE	word="i"
PE5	SortPE	sortID=2
PE7	SortPE	sortID=9
PE8	MergePE	topK=1234

Storm

Storm from Twitter

Storm



Stream, Spout, Bolt, Topology