Product Attribute Extraction Based Real-Time C2C Matching of Microblogging Messages

M.R. Mohamed Rilfi, H.M.N. Dilum Bandara, and Surangika Ranathunga
{rilfi, dilumb, surangika}@cse.mrt.ac.lk
Dept. of Computer Science and Engineering, University of Moratuwa

Motivation

- Over 70% of small business rely on social media for Consumer-to-Consumer (C2C) business opportunities.
- Sellers post product offers & buyers post their needs.
- These messages get hidden among so many others posts.
- Both buyers & sellers could benefit if such relevant messages can be detected and matched as they get posted.

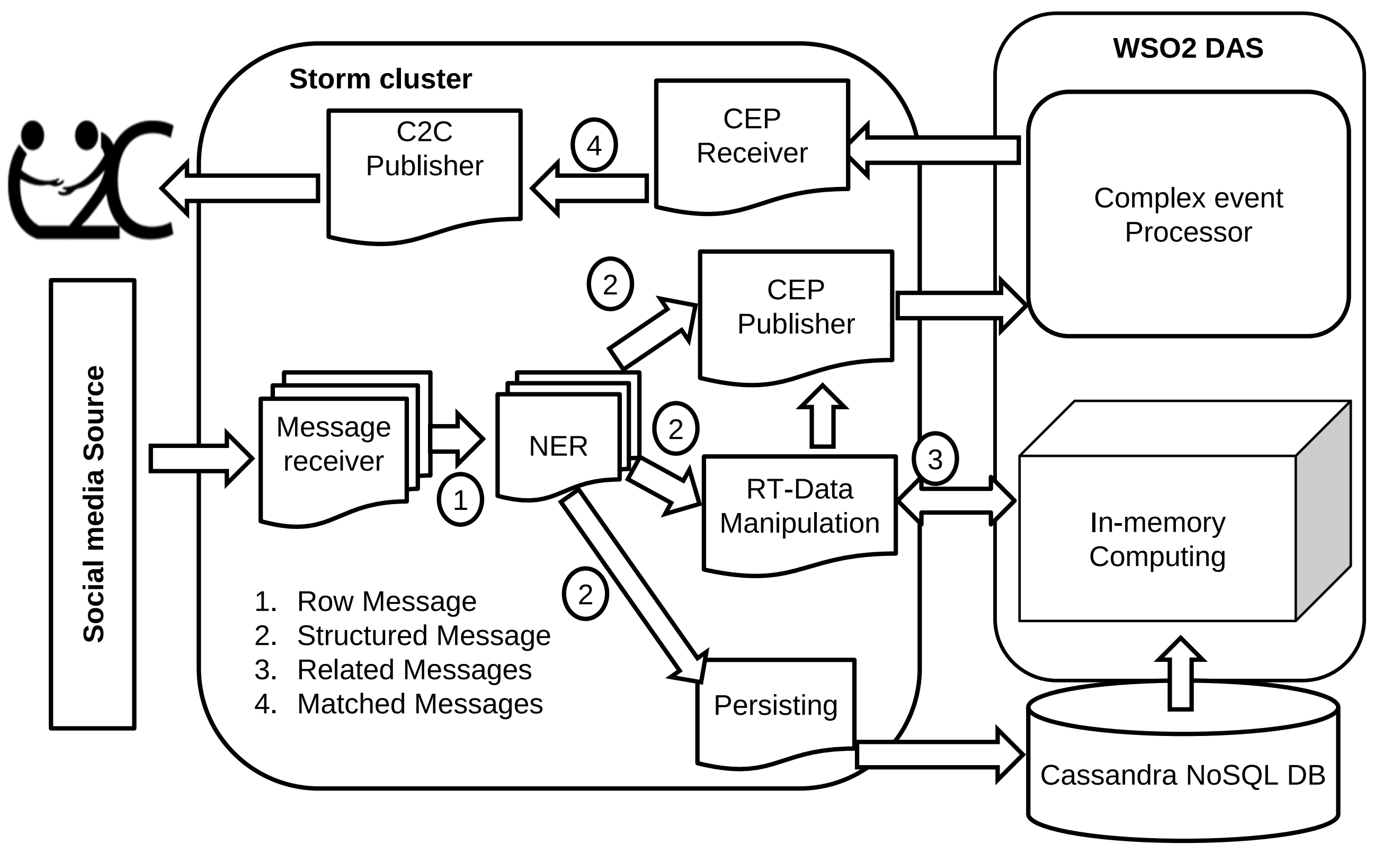
PROBLEM STATEMENT

How to develop an architecture/framework which will provide real-time C2C matching, using customers' text-based social media data?

Buyer: *#urgent need amazon kindle paper white # used around 75$*

Seller: *Kindle Paperwhite 3G, 6" HD Display, Free 3G WiFi 212 ppi. optimized font tech., 16-level gray only for $79.98*

Overall System Architecture

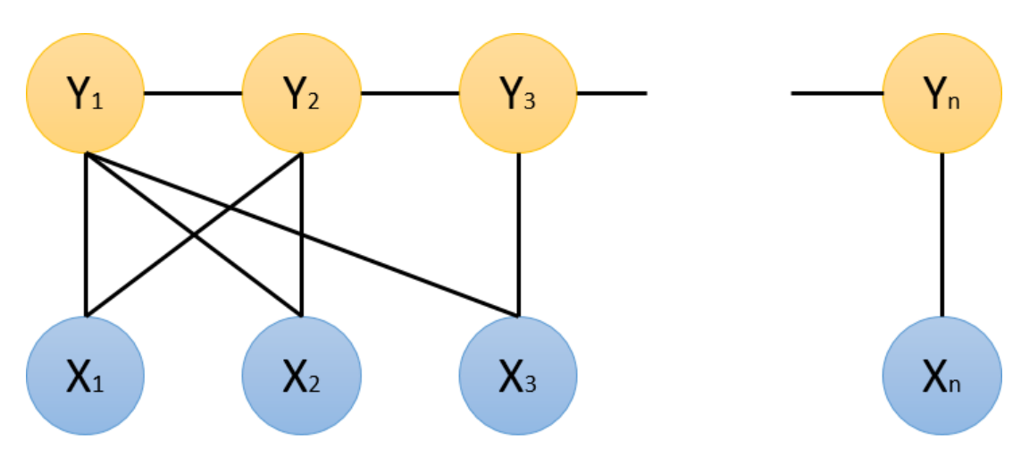


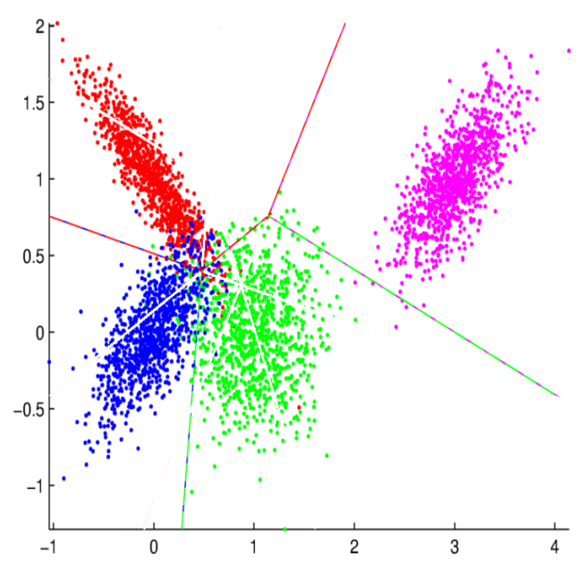
1. Row Message
2. Structured Message
3. Related Messages
4. Matched Messages

Social Media Messages

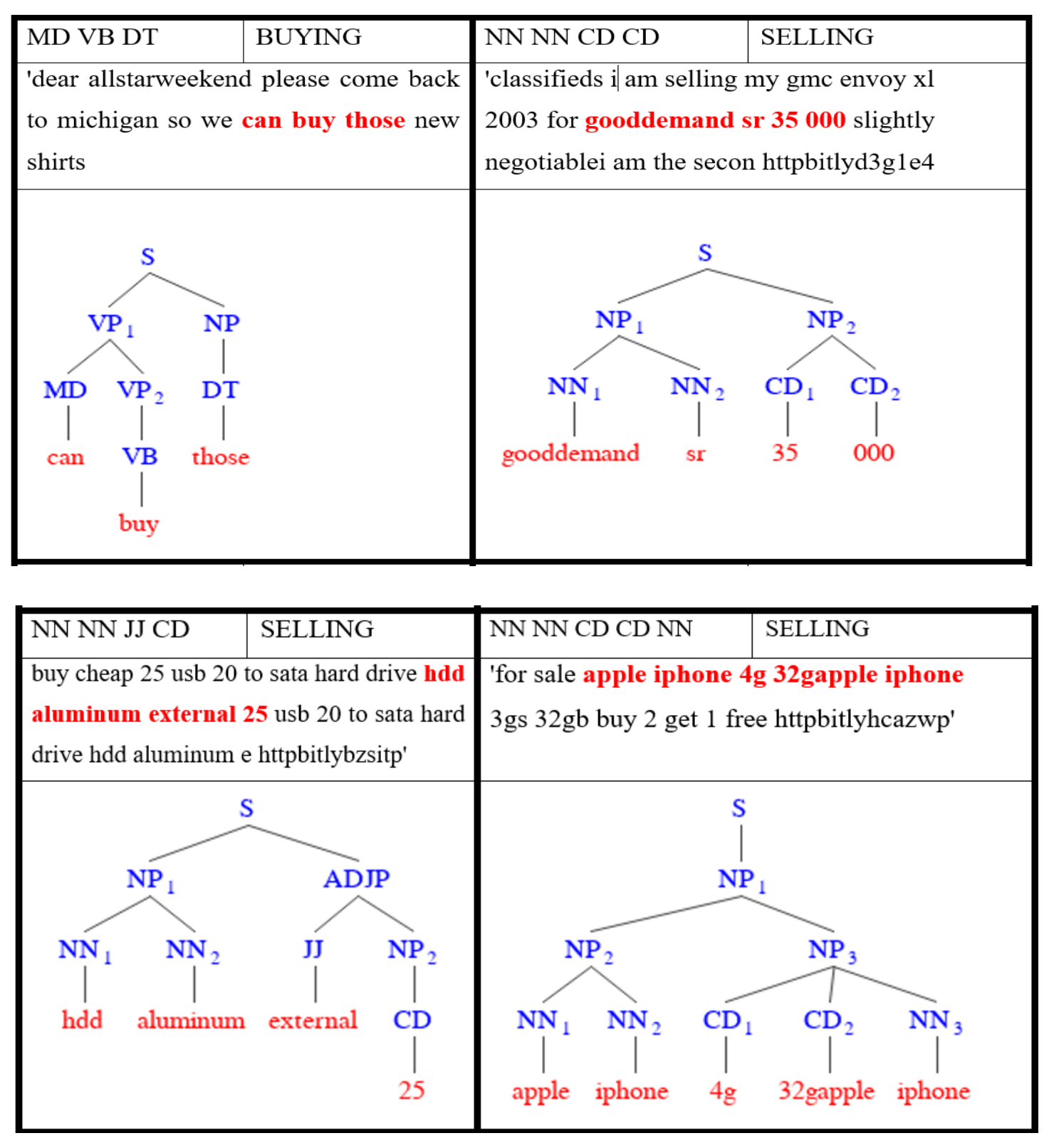
- Buying & selling related tweets for training and testing, size of 2 million
- Linked data/knowledgebase data to label our training set.
- N-Quarts(RDF) & JSON data from Amazon and web scrawling
- Amazon dataset includes 3 main product domains such as Headphones, Phones and TV

Product Attribute Extraction





Supervised method	Conditional Random Fields	Multi class logistic regression
Product attributes	Product brand	Product group
	Product name	Selling status
	Product model	



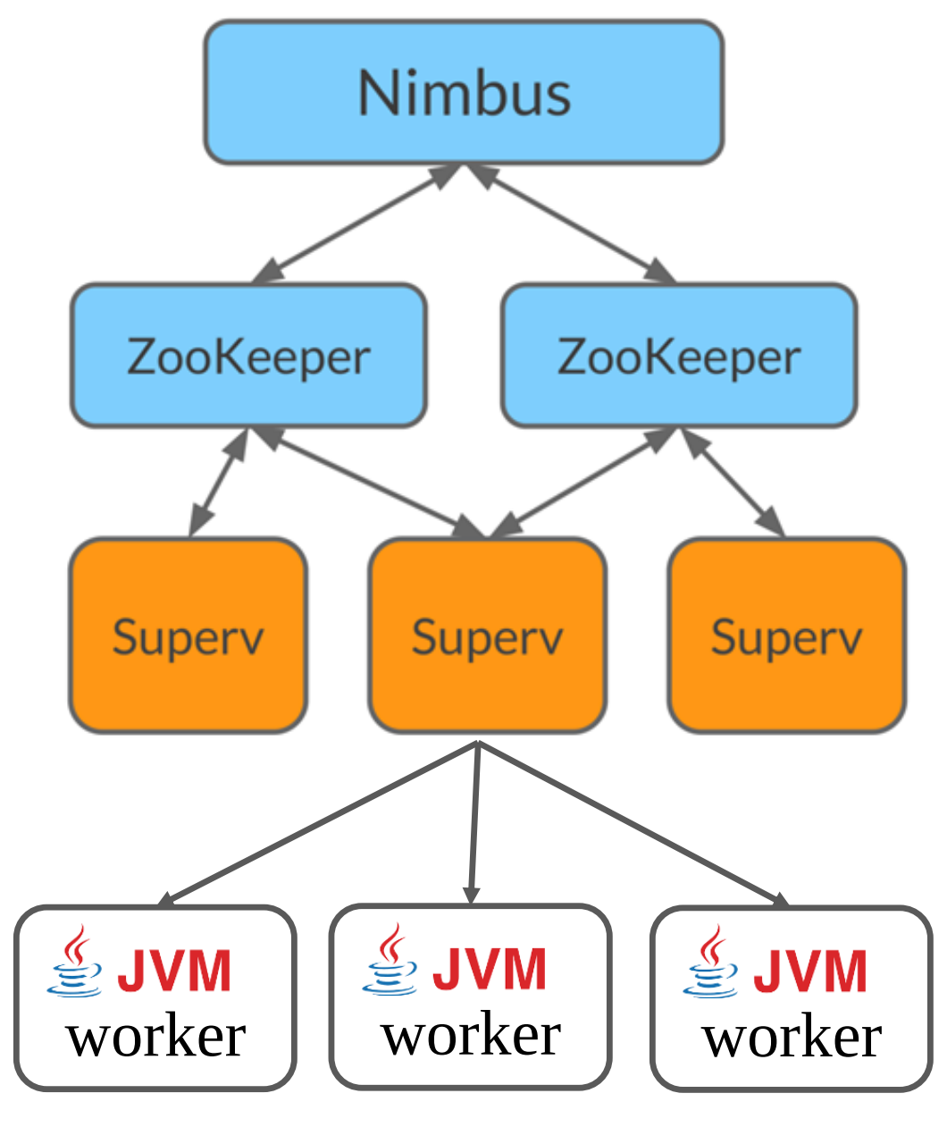
MD VB DT — BUYING
'dear allstarweekend please come back to michigan so we **can buy those** new shirts

NN NN CD CD — SELLING
'classifieds i am selling my gmc envoy xl 2003 for **gooddemand sr 35 000** slightly negotiablei am the secon httpbitlyd3g1e4

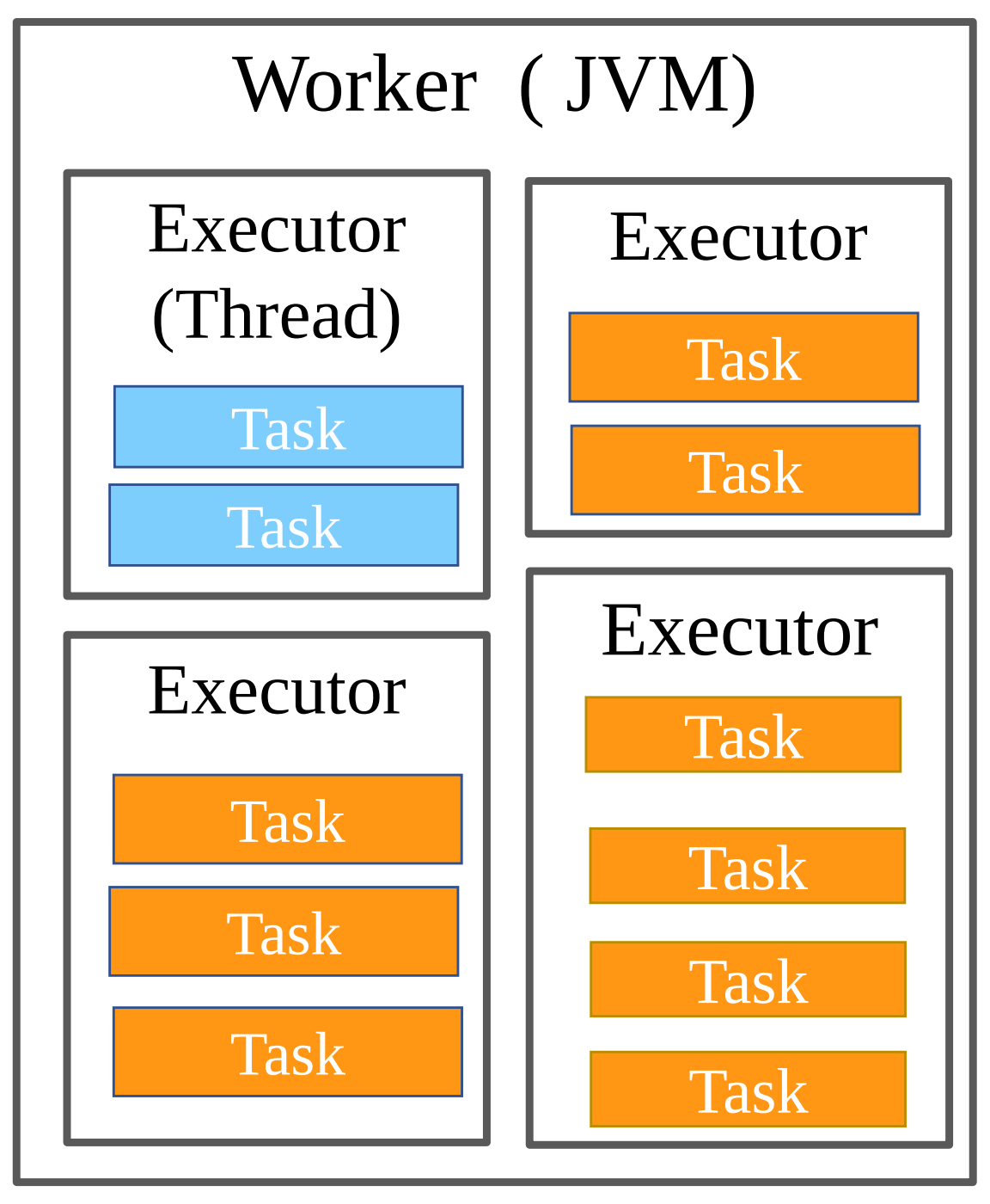
NN NN JJ CD — SELLING
buy cheap 25 usb 20 to sata hard drive **hdd aluminum external 25** usb 20 to sata hard drive hdd aluminum e httpbitlybzsitp'

NN NN CD CD NN — SELLING
'for sale **apple iphone 4g 32gapple iphone** 3gs 32gb buy 2 get 1 free httpbitlyhcazwp'

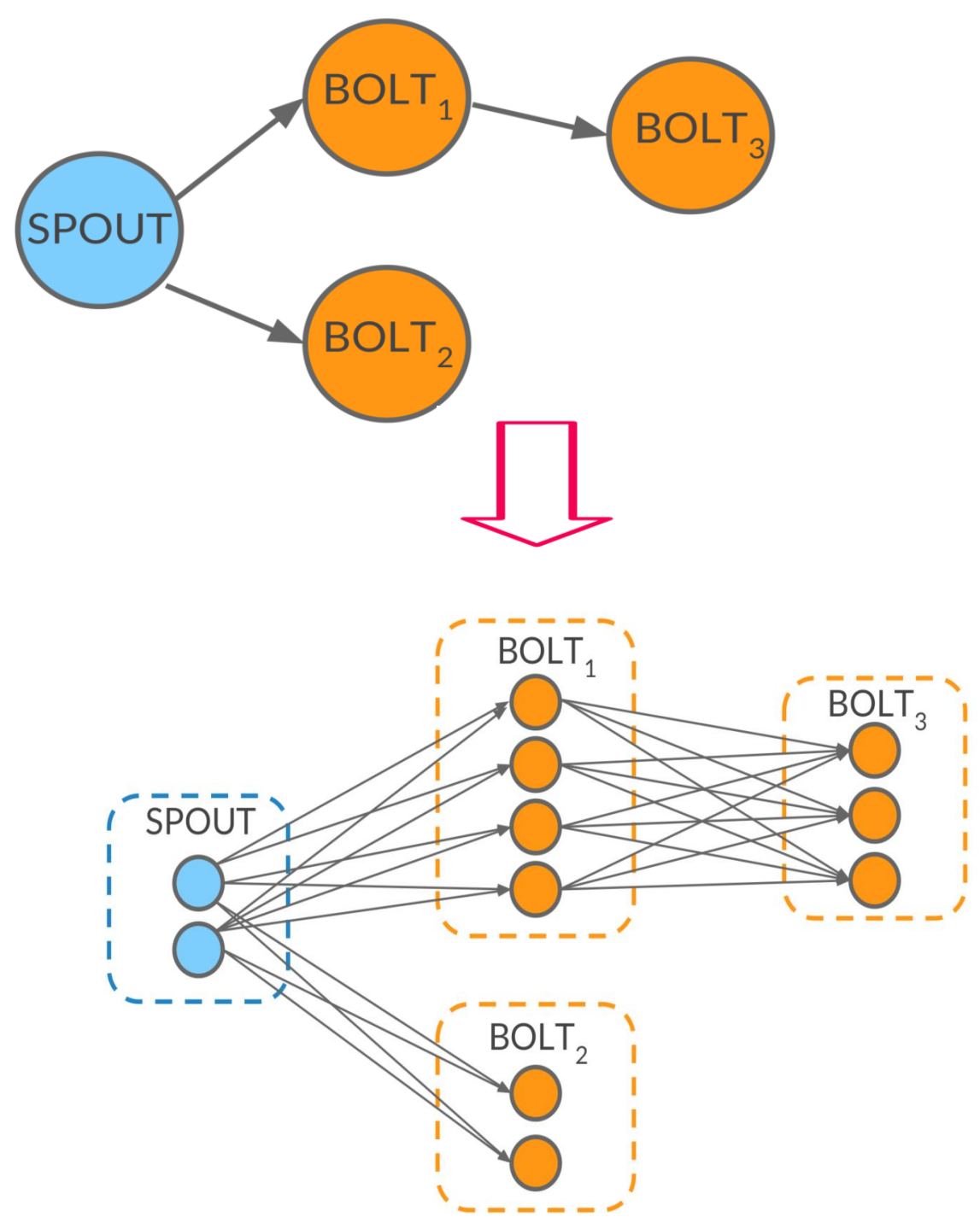
Real-time Extraction Using Distributed Stream Processing



Cluster architecture

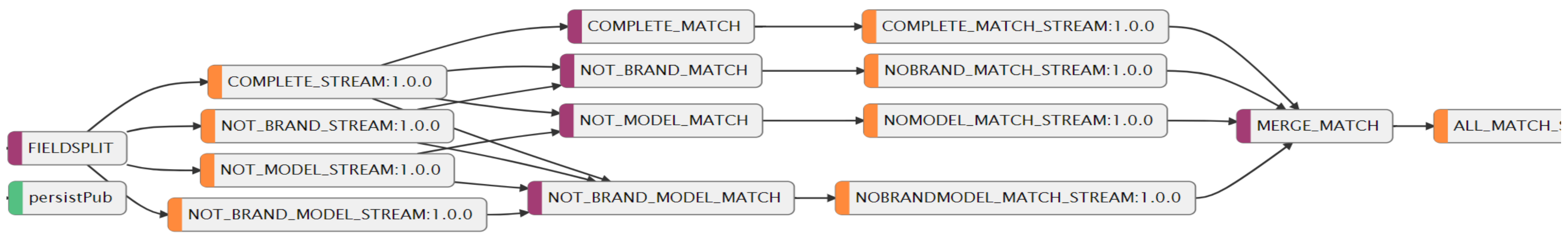


Parallelism in a single JVM

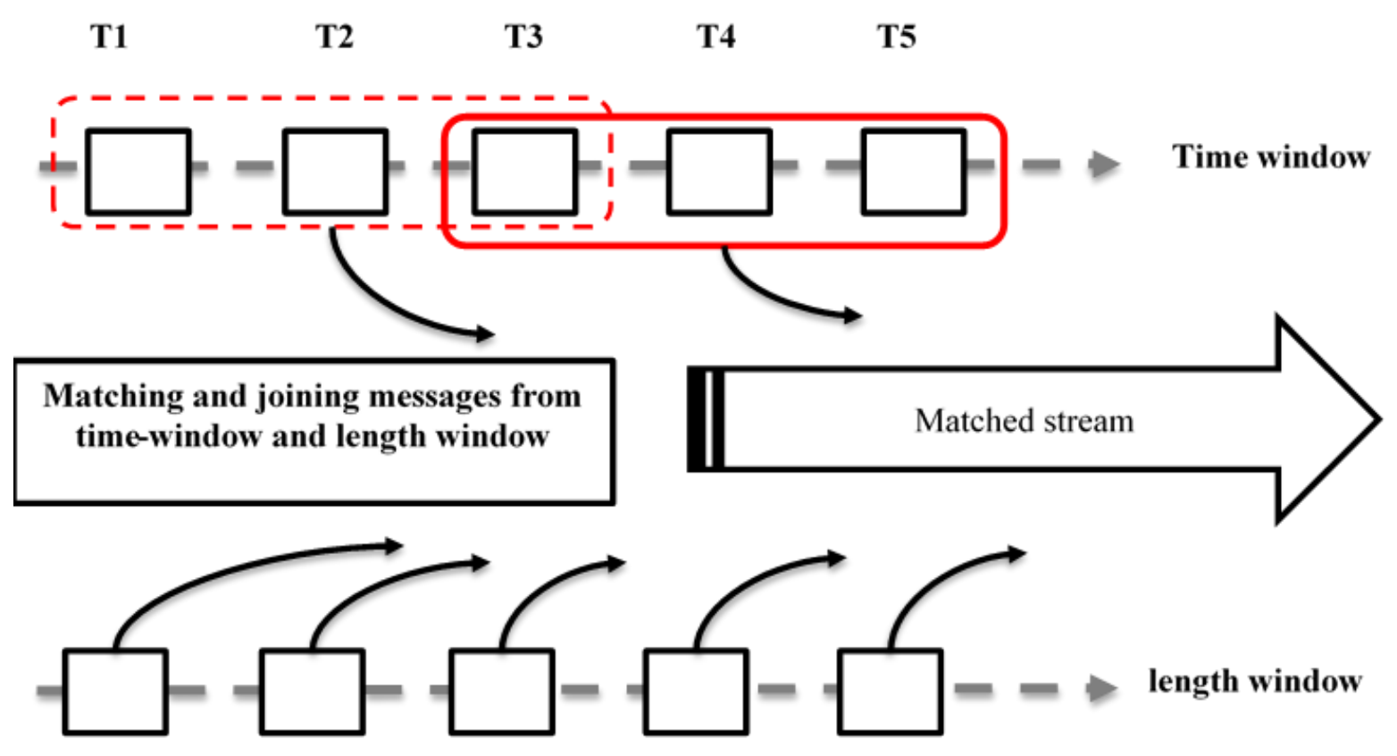


Conceptual architecture of a trivial Storm topology

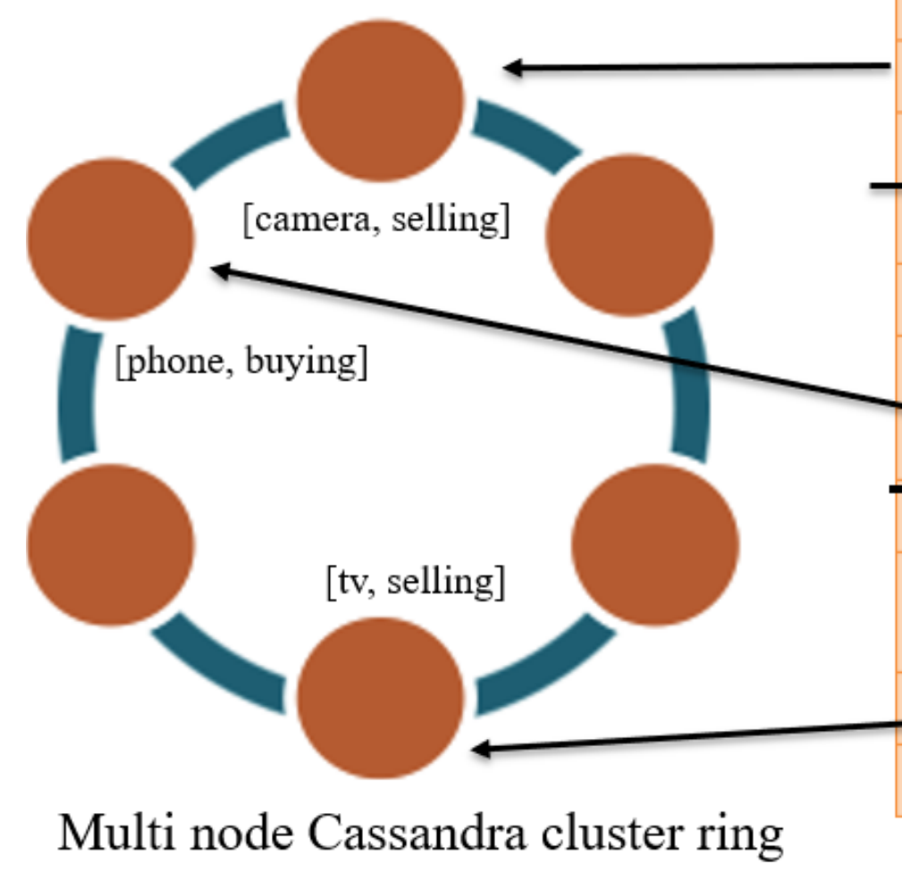
Complex Event Processing Based Product Matching



- CEP is a popular solution to process structured real-time stream
- To match buying & selling messages we used time-based matching as well as content based matching
- Time window & length window are the main 2 caching structure we using


T1
T2
T3
T4
T5
Time window
Matching and joining messages from time-window and length window
Matched stream
length window

Read Optimized Cassandra NoSQL Data Model

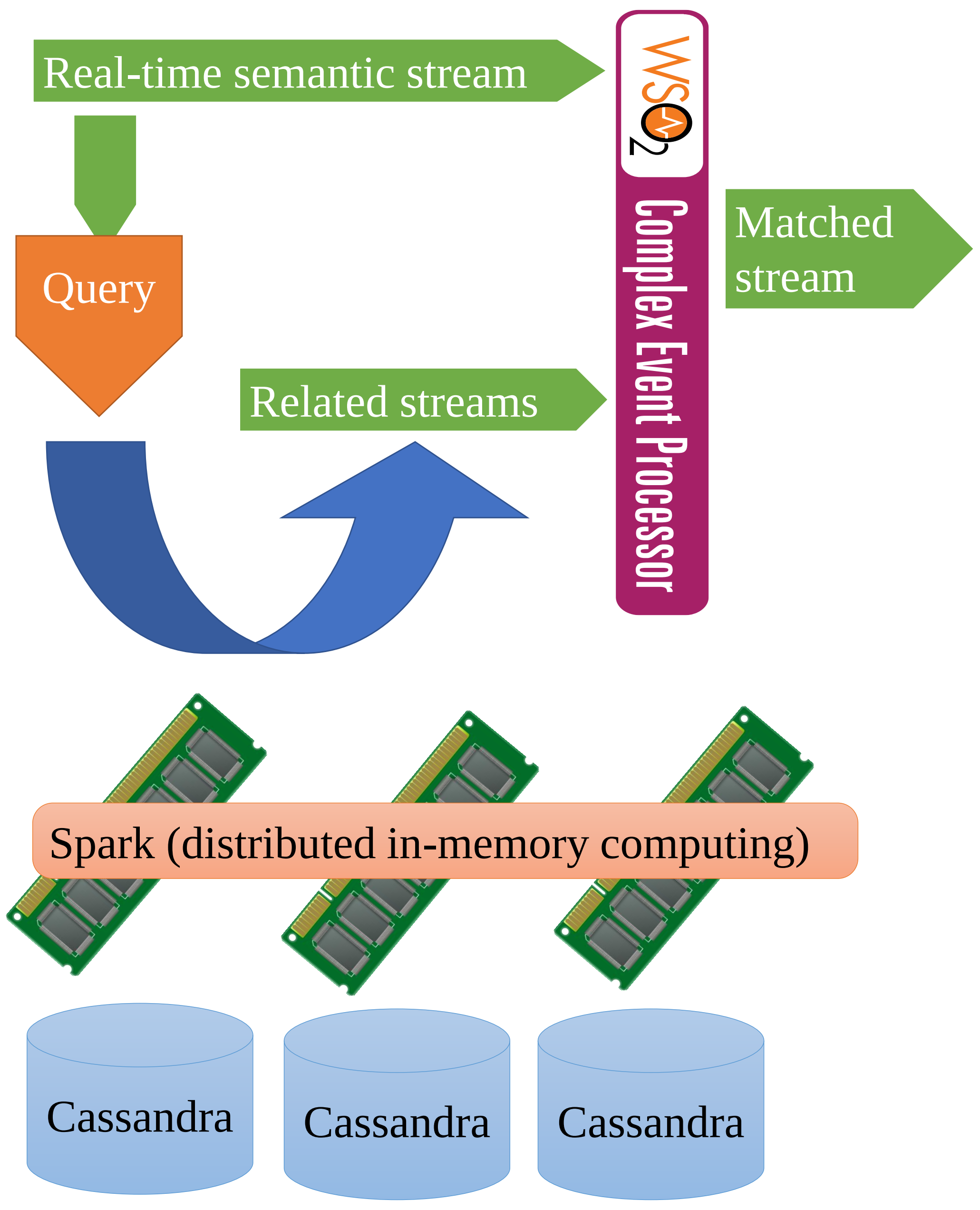


Multi node Cassandra cluster ring

Composite partition key — Clustering column

product_type	status	msg_id	brand	model	product_group
camera	selling	1	lg	g6	electronics
camera	selling	2	samsung	nx500	electronics
camera	selling	3	nikon	d800	electronics
phone	buying	4	apple	iphone 8	electronics
phone	buying	5	nokia	3310	electronics
phone	buying	6	samsung	note 8	electronics
phone	buying	7	apple	iphone 8	electronics
tv	selling	8	samsung	un55	electronics
tv	selling	9	sony	xbr-49x800e	electronics
tv	selling	10	sony		electronics
tv	selling	11	lg	60uj7700	electronics

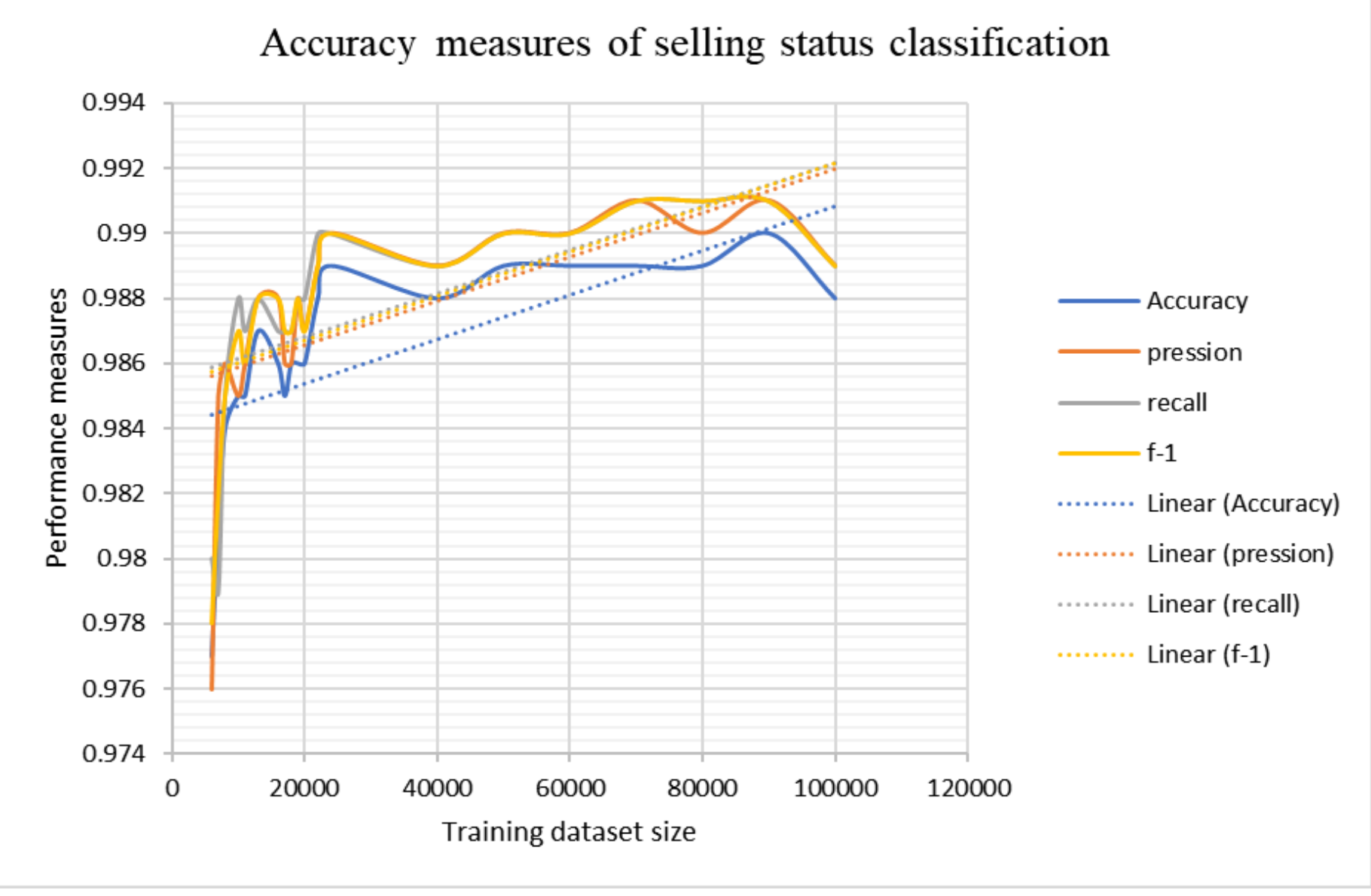
Low Latency in-Memory Processing with NoSQL

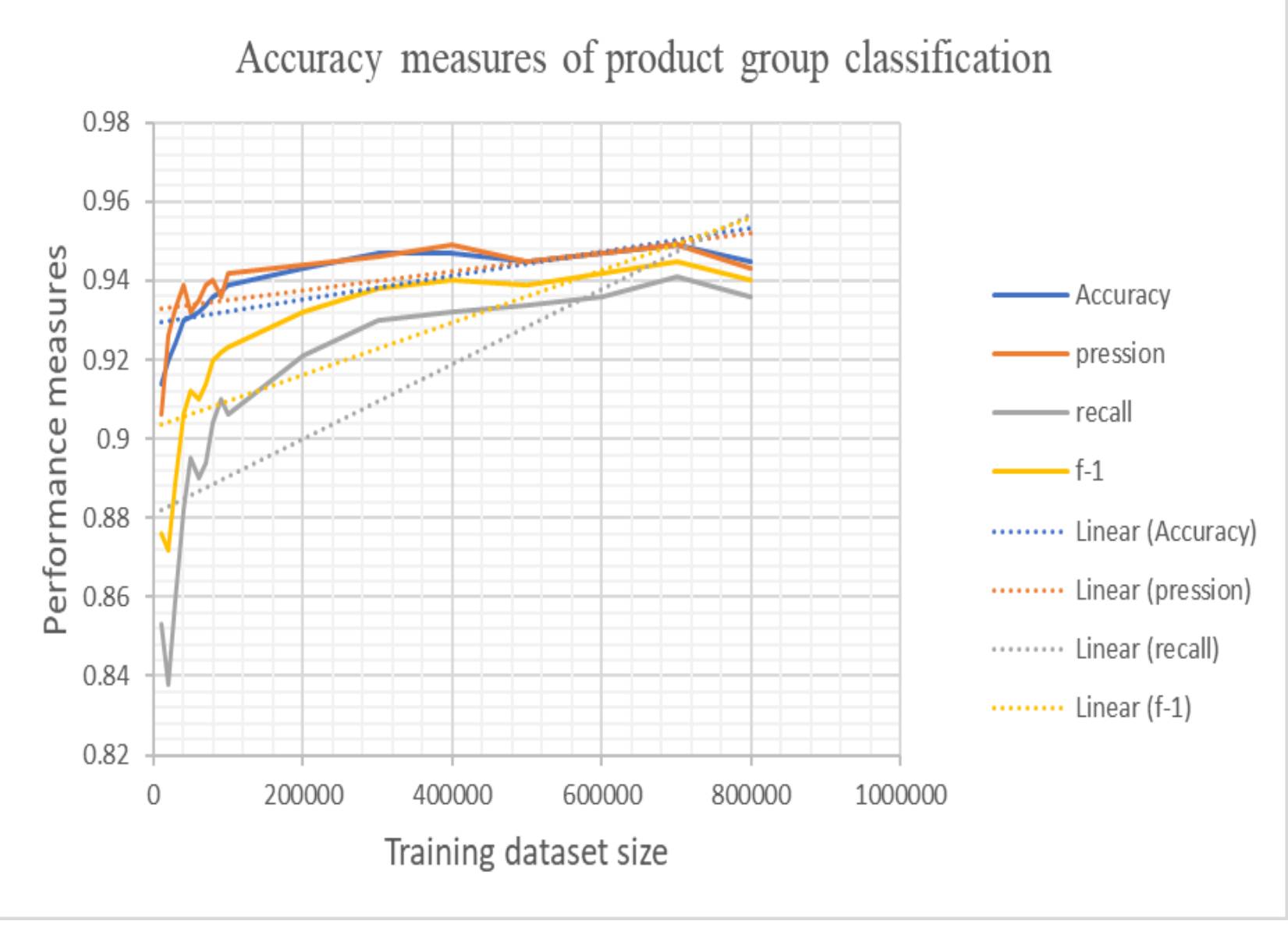


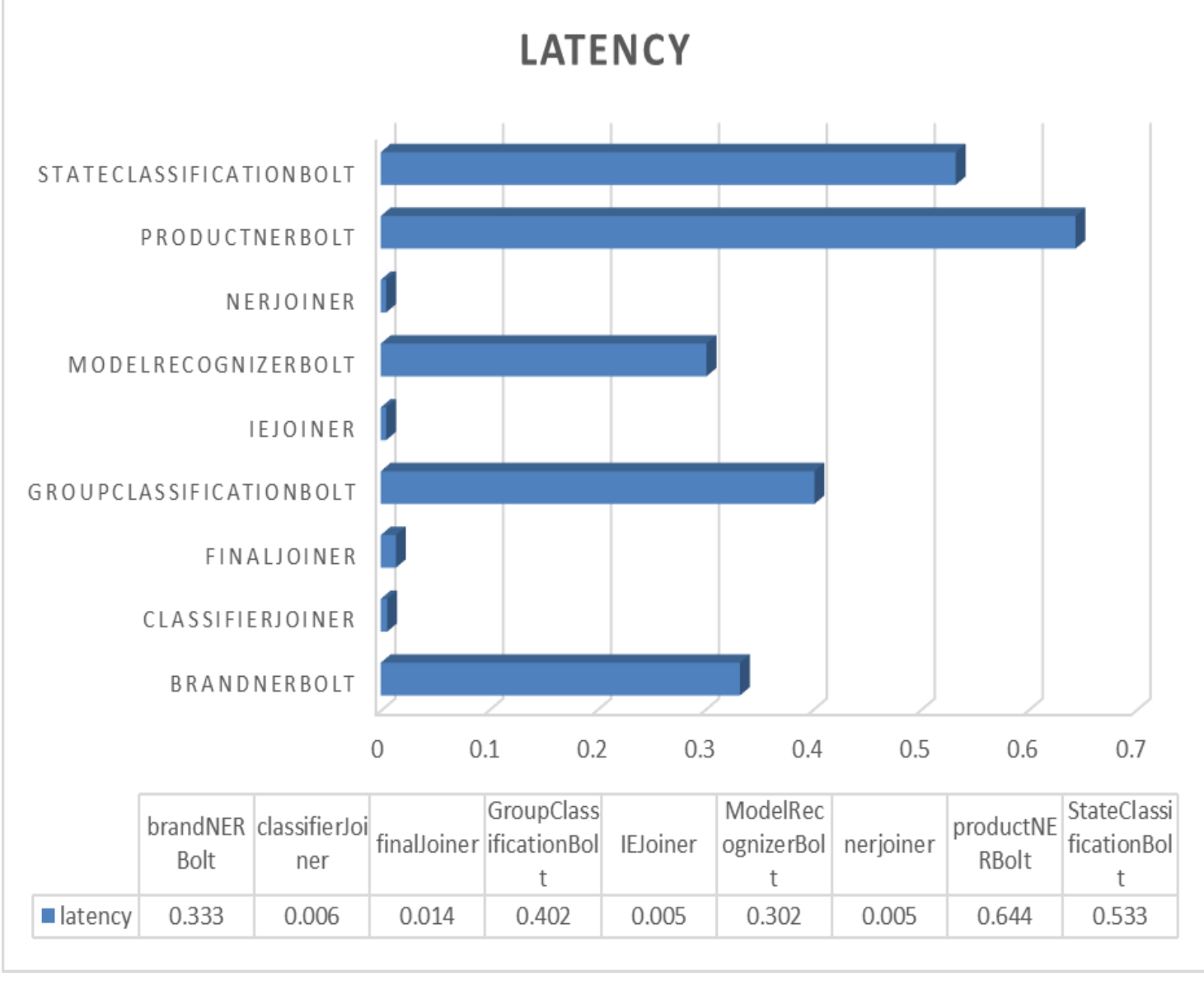
Results

ACCURACY AND PERFORMANCE MEASURES

Module Name	Accuracy	Recall	Pression	F1	Latency (MS)	Parallel Instances	Training Set Size
Brand NER	0.821	0.873	0.932	0.901	0.333	12	2,03,851
Product NER	0.84	0.904	0.922	0.913	0.644	10	
Status Classification	0.985	0.974	0.993	0.983	0.533	10	8,83,101
Product Group classification	0.948	0.96	0.944	0.952	0.402	10	9,10,951
In-memory data manipulation	-	-	-	-	5.0	-	
CEP based matching	-	-	-	-	3.6	-	







	brandNERBolt	classifierJoiner	finalJoiner	GroupClassificationBolt	IEJoiner	ModelRecognizerBolt	nerjoiner	productNERBolt	StateClassificationBolt
■ latency	0.333	0.006	0.014	0.402	0.005	0.302	0.005	0.644	0.533