Parallel Clustering of High-Dimensional Social Media Data Streams

NSF STREAM 2015 Workshop
Indianapolis, Indiana

October 27- 28, 2015

Judy Qiu

School of Informatics and Computing

Indiana University

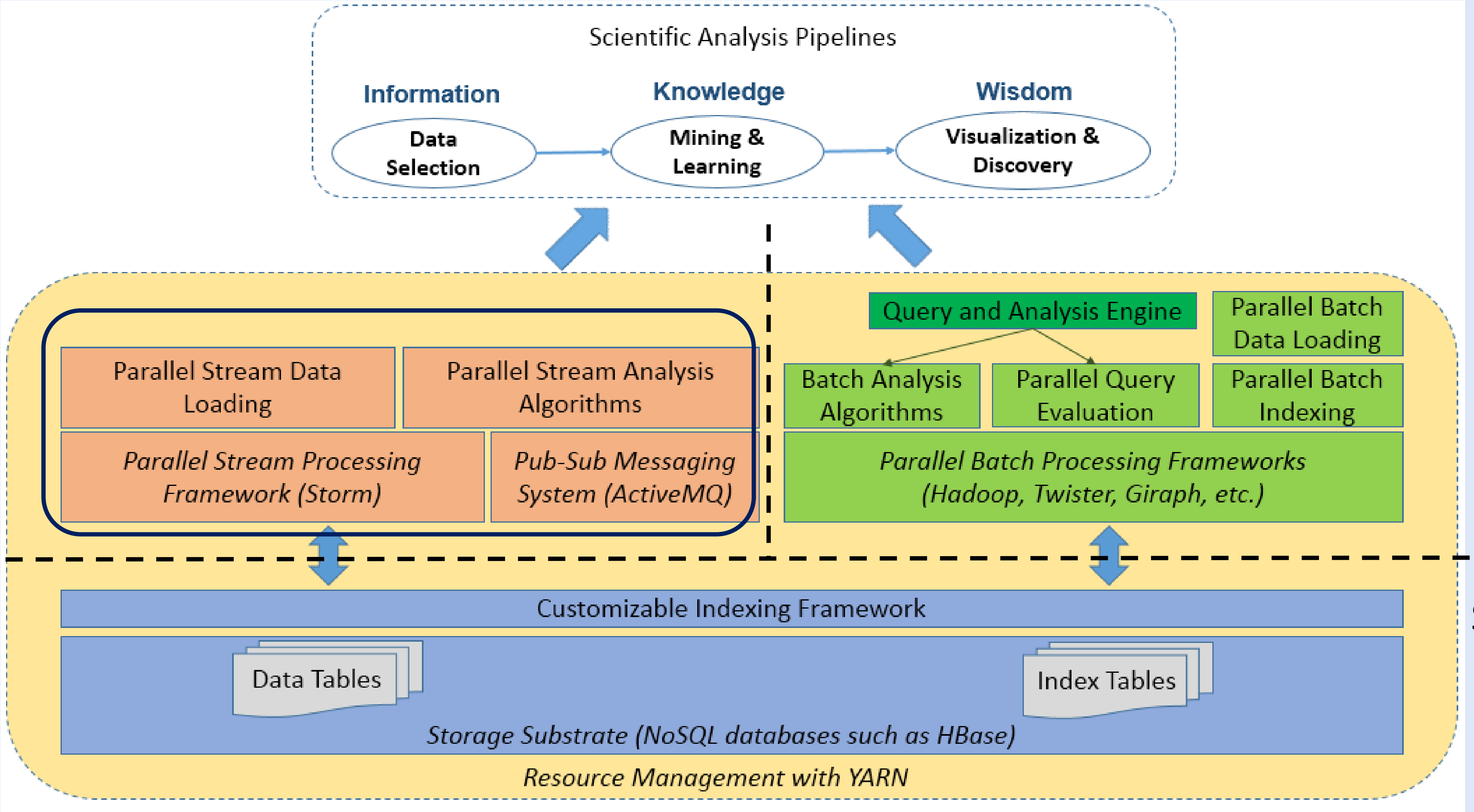
SALSA

Background

- The Convergence of Cloud, Big Data and Mobile: What could happen in 5 years?
  Big Trend: integrated batch and streaming analysis
- Example: Google DataFlow

Cloud DIKW

Streaming analysis module



Batch analysis module

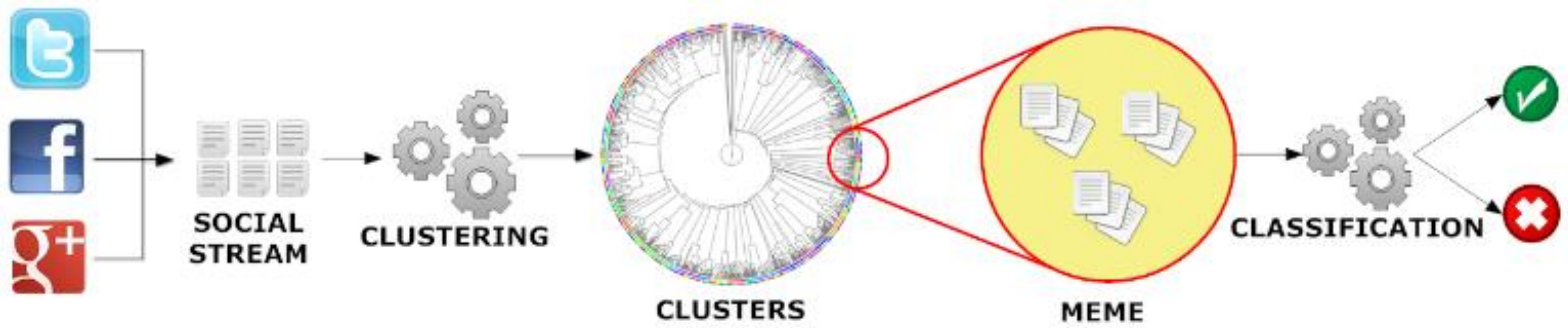
Storage substrate

- Supporting non-trivial streaming algorithms requiring global synchronization

SALSA

Parallel Tweet Online Clustering with Apache Storm

- **IU DESPIC analysis pipeline for meme clustering and classification :** Detecting Early Signatures of Persuasion in Information Cascades
- Implement with HBase + Hadoop (Batch) and HBase + Storm(Streaming) + ActiveMQ
- 2 million streaming tweets processed in 40 minutes; 35,000 clusters
- Storm Bolts coordinated by ActiveMQ to synchronize parallel cluster center updates – add loops/iterations to Apache Storm



Xiaoming Gao, Emilio Ferrara, Judy Qiu, Parallel Clustering of High-Dimensional Social Media Data Streams Proceedings of CCGrid, May 4-7, 2015

Social media data stream clustering


```
{
    "text":"RT @sengineland: My Single Best... ",
    "created_at":"Fri Apr 15 23:37:26 +0000 2011",
    "retweet_count":0,
    "id_str":"59037647649259521",
          "entities":{
        "user_mentions":[{
                "screen_name":"sengineland",
                "id_str":"1059801",
                "name":"Search Engine Land"
            }],
        "hashtags":[],
        "urls":[{
                "url":"http:\/\/selnd.com\/e2QPS1",
                "expanded_url":null
            }]},
    "user":{
        "created_at":"Sat Jan 22 18:39:46 +0000 2011",
        "friends_count":63,
        "id_str":"241622902",
        ...},
    "retweeted_status":{
        "text":"My Single Best... ",
        "created_at":"Fri Apr 15 21:40:10 +0000 2011",
        "id_str":"59008136320786432",
        ...},
    ...
}
```


- Group social messages sharing similar social meaning
- Useful in meme detection, event detection, social bots detection, etc.

SALSA

Social media data stream clustering

- Recent progress in learning **data representations** and **similarity metrics**
- High-quality clusters: leverage both textual and network information (high-dimensional vectors)
- Expensive similarity computation: 43.4 hours to cluster 1 hour's worth of data with sequential algorithm
- Goal: meet real-time constraint through parallelization
- Challenge: efficient global synchronization in DAG-oriented parallel processing frameworks

SALSA

Sequential algorithm for clustering tweet stream

- Online K-Means with sliding time window and outlier detection
- Group tweets as **protomemes**: hashtags, mentions, URLs, and phrases.
- Cluster protomemes using similarity measurement:

- Common **user** similarity: $S_u(P_i, P_j) = \frac{\sum_{u \in U_i \cap U_j} P_{iu} P_{ju}}{\sqrt{\sum_{u \in U_i} P_{iu}^2} \sqrt{\sum_{u \in U_j} P_{ju}^2}}$

- Common **tweet ID** similarity: $S_t(P_i, P_j) = \frac{|P_i \cap P_j|}{\sqrt{|P_i|}\sqrt{|P_j|}}.$

- **Content** similarity: $S_c(P_i, P_j) = \frac{\sum_{w \in W_i \cap W_j} P_{iw} P_{jw}}{\sqrt{\sum_{w \in W_i} P_{iw}^2} \sqrt{\sum_{w \in W_j} P_{jw}^2}}$

- **Diffusion** similarity: $S_d(P_i, P_j) = \frac{|N_i \cap N_j|}{\sqrt{|N_i|}\sqrt{|N_j|}}$ $N_\ell = U_\ell \cup M_\ell \cup R_\ell$ (Posting + mentioned + retweeting)

- Combinations: $MAX(P_i, P_j) = \max_k \{S_k(P_i, P_j)\}$ $\mathcal{L}(P_i, P_j) = \sum_k \omega_k S_k(P_i, P_j)$

Sequential Algorithm for Clustering Tweet Stream

- Online (streaming) K-Means clustering algorithm with *sliding time window* and *outlier detection*
- Group tweets in a time window as **protomemes**:
  - Label protomemes (points in space to be clustered) by "markers", which are *Hashtags*, *User mentions*, *URLs*, and *phrases*
  - A phrase is defined as the textual content of a tweet that remains after removing the hashtags, mentions, URLs, and after stopping and stemming
    - Number of tweets in a *protomeme* : Min: 1, Max :206, Average 1.33
- Note a given tweet can be in more than one protomeme
  - One tweet on average appears in 2.37 protomemes
  - Number of protomemes is 1.8 times number of tweets

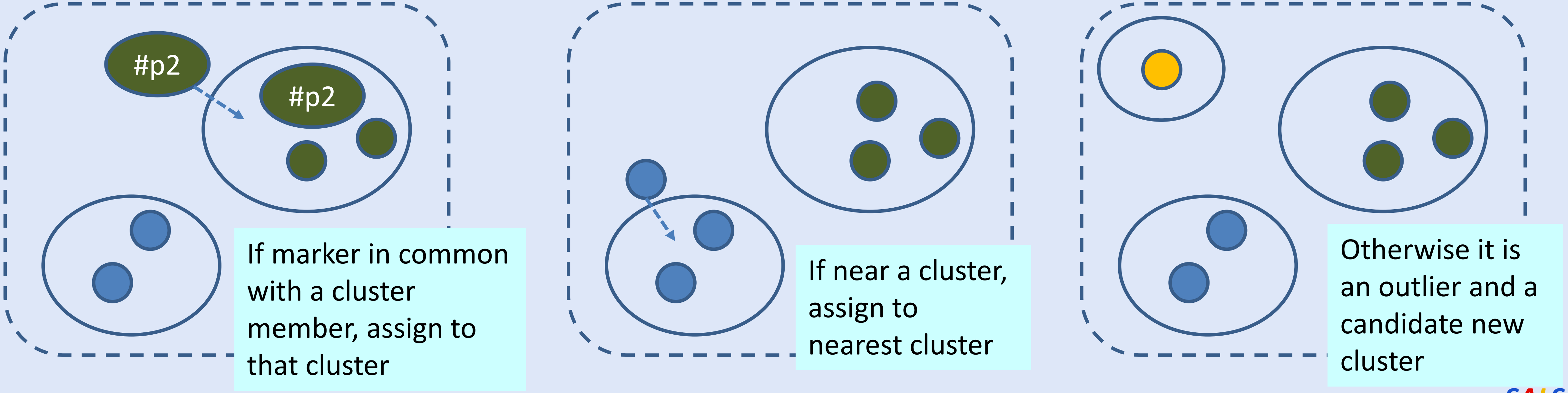
Online K-Means clustering

(1) Slide time window by one time step

(2) Delete old protomemes out of time window from their clusters

(3) Generate protomemes for tweets in this step

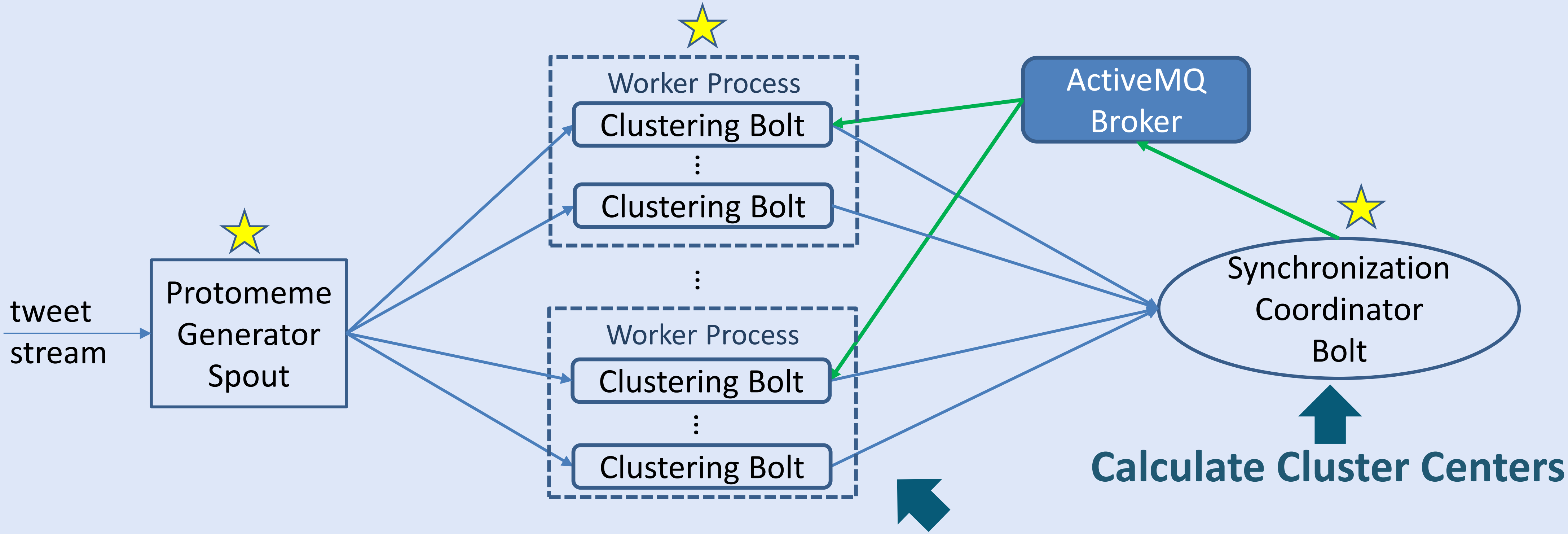
(4) For each new protomeme classify in old or new cluster (outlier)



SALSA

Parallelization with Storm – Challenges

DAG organization of parallel workers: hard to synchronize cluster information



Parallelize Similarity Calculation

☆ Synchronization initiation methods:

- Spout initiation by broadcasting INIT message
- Clustering bolt initiation by local counting
- Sync coordinator initiation by global counting (of #protomemes)

Suffer from variation of processing speed

Scalability Comparison

Full-centroids synchronization

Number of clustering bolts	Total processing time (sec)	Compute time / sync time	Sync time per batch (sec)	Avg. length of sync message
3	67603	30.3	6.71	22,113,520
6	35207	15.1	6.71	21,595,499
12	19295	7.0	7.32	22,066,473
24	11341	3.2	8.24	22,319,413
48	7395	1.5	9.15	21,489,950
96	6965	0.7	12.93	21,536,799

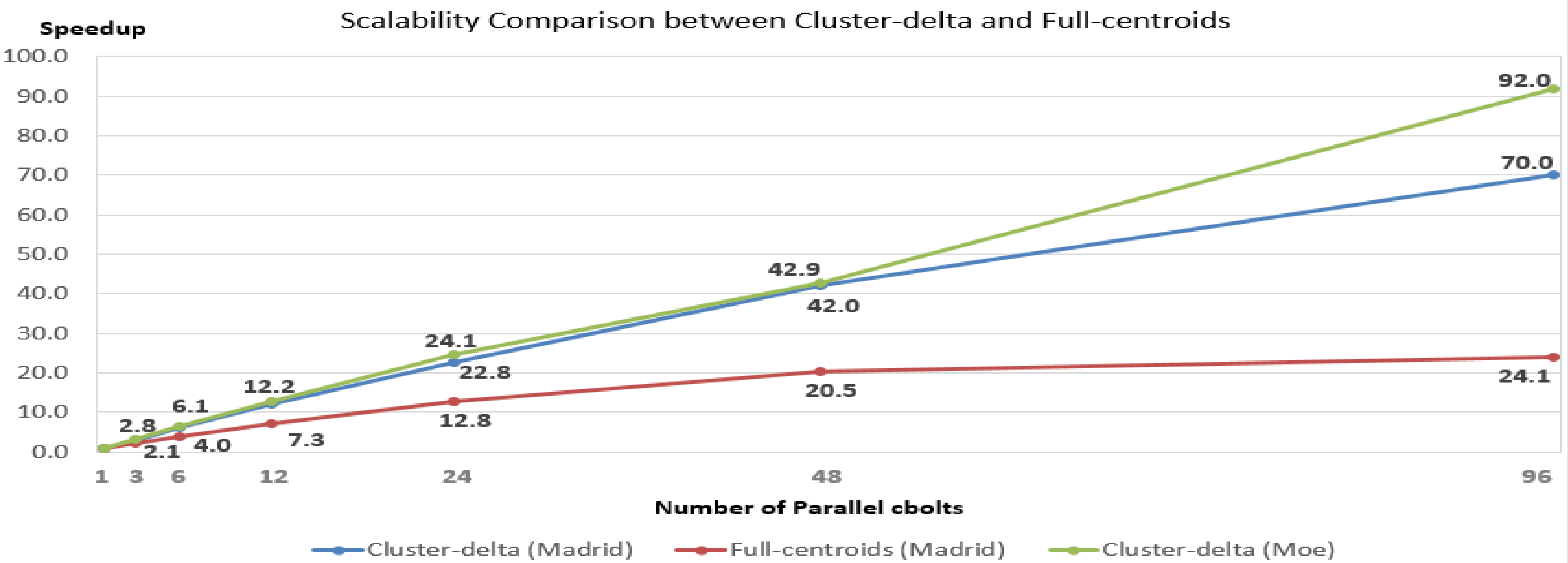
Cluster-delta synchronization

Number of clustering bolts	Total processing time (sec)	Compute time / sync time	Sync time per batch (sec)	Avg. length of sync message
3	50381	252.6	0.62	2,525,896
6	22949	96.4	0.73	2,529,779
12	11560	42.2	0.81	2,532,349
24	6221	21.7	0.81	2,544,095
48	3490	8.4	1.08	2,559,221
96	2494	2.5	2.17	2,590,857

Reduce synchronization overhead by sending incremental changes to the centroid vector.

SALSA

Parallel Tweet Clustering with Storm



- Speedup on up to 96 bolts on two clusters, Moe and Madrid
- Red curve is old online K-means algorithm; green and blue are the new algorithm
- Full Twitter – 1000 way parallelism (expected)

SALSA

Acknowledgements

Xiaoming Gao

Prof. Filippo Menczer & CNETS
Complex Networks and Systems

SALSA   HPC Group
School of Informatics and Computing
Indiana University