Models and Issues in Data Stream Systems

Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, Jennifer Widom

Presented by Christian Valdemar Mathiesen
cmath@cs.brown.edu
March 9, 2015

STREAM*

*STanford StREam DatA Manager

STREAM

- Query language
- Query processing
- Conclusion

Query language

"In the STREAM project, we have chosen to use a modified version of SQL as the query interface to the system […]. SQL is a well-known language with a large user population."

```
SELECT   AVG(V.minutes)
FROM     (SELECT S.minutes
          FROM Calls S, Customers T
          WHERE S.customer_id = T.customer_id
          AND T.tier = 'Gold')
          V [ROWS 1000 PRECEDING]
```

VS.

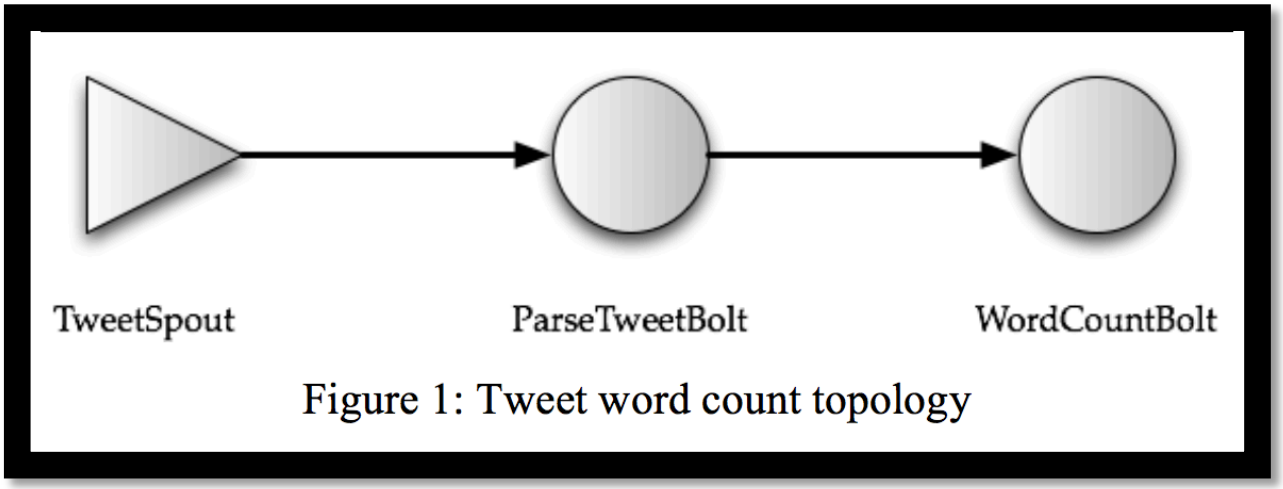


Figure 1: Tweet word count topology

Source: "Storm @Twitter", Toshniwal et al.

Which is easier to understand?

STREAM

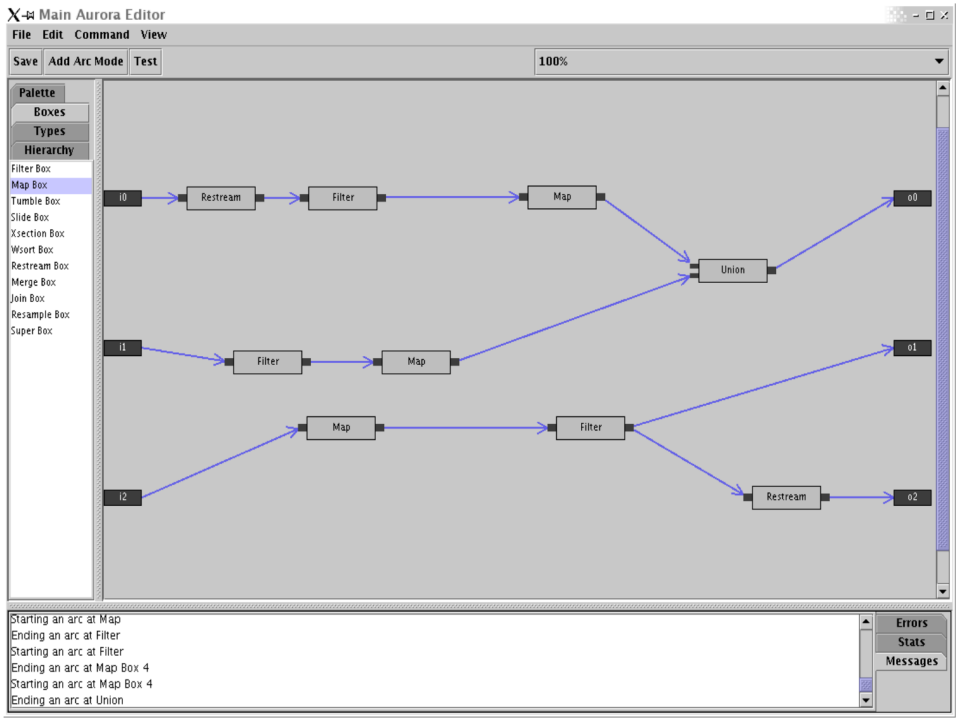
```
SELECT lowerImages.*, higherImages.*
FROM globalrank AS lowerImages, globalrank AS higherImages
WHERE lowerImages.rank < higherImages.rank
AND lowerImages.image_id = (
    SELECT A.image_id AS lower_image_other (
    SELECT lowerImages.image_id AS lower_image,
           max(higherImages.image_id) AS higher_image
    FROM global_rank AS lowerImages, global_rank AS higherImages
    WHERE  lowerImages.rank  <  higherImages.rank

    AND 1  NOT IN (select 1 from ranked_up where
        lowerImages.image_id = ranked_up.image_id
        AND ranked_up.user_id = $user_id
        AND ranked_up.created_at > DATE_SUB(NOW(), INTERVAL 1 DAY))

    AND 1 NOT IN (
        SELECT 1 from matchups where user_id = $userId
                AND lower_image_id = lowerImages.image_id
                AND higher_image_id = higherImages.image_id
                UNION
                SELECT 1 from matchups where user_id = $user_id
                AND lower_image_id = higherImages.image_id
                AND higher_image_id = lowerImages.image_id
    )
) A
AND NOT EXISTS (
    SELECT * FROM matchups
    WHERE user_id = $user_id
    AND ((image_id1 = lowerImages.image_id AND image_id2 = higherImages.image_id)
        OR (image_id2 = lowerImages.image_id AND image_id1 = higherImages.image_id))
)
AND higherImages.image_id NOT IN (
    SELECT image_id FROM rankedup
    WHERE created_at < DATE_ADD(NOW(), INTERVAL 1 DAY)
    AND USER_ID <> $user_id
)
ORDER BY higherImages.rank
```

*

Aurora



**

*Source: http://stackoverflow.com/questions/6564601/sql-query-with-complex-subqueries

** Source: The Aurora and Borealis Stream Processing Engines, Cetintemel et al.

Timestamps

*"Formally we say that a data stream consists of a set of (tuple, timestamp) pairs[...] — **all that is required is that [the timestamp] comes from a totally ordered domain** with a distance metric."*

Timestamps

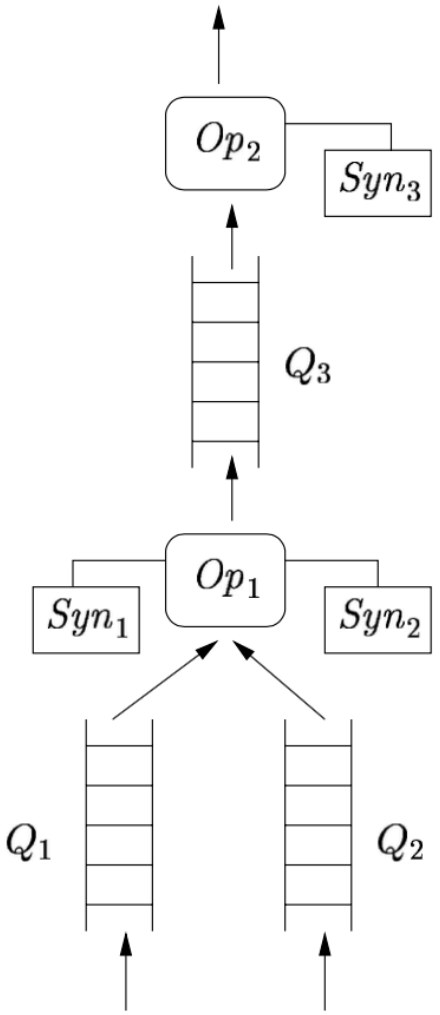
What if tuples arrive from multiple sources?

In other words, how do we guarantee a **totally ordered domain**?

Query processing

Paper uses same notation for queries and queues!?

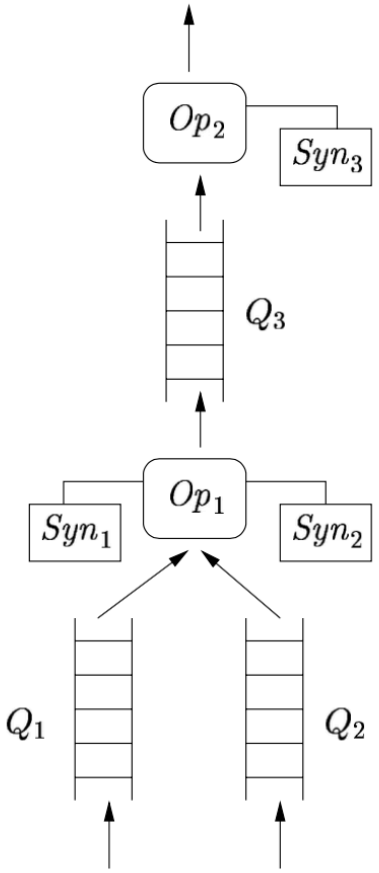
Q_1: SELECT `notifyoperator`(sum(len))
FROM B
GROUP BY `getminute`(time)
HAVING sum(len) $>$ t



Query processing

How are query plans generated?

How does the system scale (i.e. it only has one central scheduler)?



Conclusion

- Paper presents a series of relevant issues for OLTP systems
- STREAM tries to solve these issues, but reasoning behind design decisions are sometimes unclear
- Algorithmic issues should be put in separate paper