Distributed Data Management
Summer Semester 2013
TU Kaiserslautern

Dr.-Ing. Sebastian Michel

smichel@mmci.uni-saarland.de
(DISTRIBUTED) DATA STREAM PROCESSING
Data Stream Management vs.
Traditional Data Management

• **Data is moving!** Continuously generated (assumed infinite!)
• At high pace
• **Queries are (mainly) continuous** (aka. standing). Registered once, observed “forever”.
• Answer to queries in (near) **real-time required** (often)
• Probabilistic methods for efficiency or considering only part of the stream (**sliding window**)
### DBMS vs. DSMS

<table>
<thead>
<tr>
<th>Database management system (DBMS)</th>
<th>Data stream management system (DSMS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistent data (relations)</td>
<td>volatile data streams</td>
</tr>
<tr>
<td>Random access</td>
<td>Sequential access</td>
</tr>
<tr>
<td>One-time queries</td>
<td>Continuous queries</td>
</tr>
<tr>
<td>(theoretically) unlimited secondary storage</td>
<td>limited main memory</td>
</tr>
<tr>
<td>Only the current state is relevant</td>
<td>Consideration of the order of the input</td>
</tr>
<tr>
<td>relatively low update rate</td>
<td>potentially extremely high update rate</td>
</tr>
<tr>
<td>Little or no time requirements</td>
<td>Real-time requirements</td>
</tr>
<tr>
<td>Assumes exact data</td>
<td>Assumes outdated/inaccurate data</td>
</tr>
<tr>
<td>Plannable query processing</td>
<td>Variable data arrival and data</td>
</tr>
<tr>
<td></td>
<td>characteristics</td>
</tr>
</tbody>
</table>


Distributed Data Management, SoSe 2013, S. Michel
Data Stream Model

• Stream of data items is unbounded (available memory is not)
• No way to store entire stream (how could we, its (probably) not ending)
• To compute query results, need to devise algorithm with little memory consumption
Overview of Data Stream Topics

• **Synopses:**
  – concise representations of stream content
  – tailored to tasks, e.g., counting distinct elements
  – usually not exact, but approximations (estimators) of true values.

• **Windows:**
  – focus of certain recent subset of data
  – computation of functions/joins over window(s) content
  – Think: SQL over stream windows (ranges)
SYNOPSIS (ESTIMATORS)
Counting Occurrences

• Consider a stream of elements $a_i$
  ..., $a_2$, $a_{84}$, $a_{41}$, $a_2$, $a_{77}$, $a_{231}$, $a_2$, $a_4$, $a_{54}$, ...

• How often does $a_2$ occur?

• How to implement?

• Keep counter for each id

• Required space $\#ids (=N)$

• Not feasible if $N$ is very large
Probabilistic Counting: Count-Min Sketch

• Keep 2-dim array \((h, r)\)
• \(h\) hash functions* that map to range 0...(r-1)

\[
\begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 & 5 \\
\h_1 & & & & & \\
\h_2 & & & & & \\
\h_3 & & & & & \\
\h_4 & & & & & \\
\end{array}
\]

• Arriving item \(a\)
• For each \(j\): \(\text{array}[j, \h_j(a)]++\)

Count-Min Sketch: Counting

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>h₁</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>h₂</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>h₃</td>
<td>8</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>h₄</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>

- How often did we see item a?
- \( h₁(a) = 4, \ h₂(a)=5, \ h₃(a)=0, \ h₄(a)=2 \)
- Take minimum of the corresponding values in the 2-d array. Here: 4
- Estimate is never underestimating
- Overestimation probabilistically bounded
Unbiased vs. Biased Estimators

• Given a real number \( n \) and an estimator of it, denoted as \( \hat{n} \)

• E.g., number of distinct elements in a set \( S \)

• \( \hat{n} \) is called an unbiased estimator of \( E[\hat{n}] = n \)

• and biased otherwise, in which case

  \[ \text{Bias}[\hat{n}] = E[\hat{n}] - n \]
Counting Distinct Elements

• Consider a stream of elements $a_i$
  
  $\ldots, a_2, a_{84}, a_{41}, a_2, a_{77}, a_{231}, a_2, a_4, a_{54}, \ldots$

• How to compute/estimate the number of distinct elements observed?
Usability

• Streams (one pass, little memory footprint)

• Distributed systems: compact data exchange (recall Bloom filter)

• Sketches for partial data can be merged for global view
Flajolet Martin (FM) Sketch (aka. Hash Sketch)

• Proposed originally by Flajolet and Martin in 1985
• Allocate a bitvector B of size \( m = \log(N) \)
• Hash items to bitvector positions according to a geometric distribution:
  
  – Hash each item \( i \) to a \( m \)-bit number \( h(i) \)
  – Compute position \( k \) of the least-significant “1” of \( h(i) \)
  – Set the bit \( B[k] \) to “1”

\[
S: 17, 5, 19, 211, 17, 5, 31
\]

\[
h(17) = 010100 \text{ then least-sig. } 1 \text{ bit } = 3
\]

\[
h(5) = 000101 \text{ then least-sig. } 1 \text{ bit } = 1
\]
FM-Sketch: Estimator

• In the end B might look like this
  B: 111010
• Get then position t of left-most “0” bit of B
• Count-Distinct Estimate of real distinct number n:
  \[ \hat{n} = 2^t / 0.7735 \]
  Here: with \( t = 4 \): \[ \hat{n} = 2^4 / 0.7735 \approx 20.685 \]
• Improvement:
  – If you use more bitmaps and compute an average position \( t \), you can improve count-distinct estimate

Note: Be careful with left-most bit; depends on interpretation of bits
FM Sketch: Intuition/Idea

- $B[0]$ is set approximately $n/2$ times
- $B[1]$ is set approximately $n/4$ times
- $B[i] = 0$ if $i \gg \log_2(n)$
- $B[i] = 1$ if $i \ll \log_2(n)$
- “Mix” of 1s and 0s around $i \approx \log_2(n)$

- Use left-most zero at indicator for $\log_2(n)$:
  \[ n \approx 2 \text{ position of left most zero bit} \]
FM-Sketch: Union

• Given: two multisets $S$ and $T$ and their sketches $B_T$ and $B_S$ of size $m$

• Then:

The sketch $B = B_S \lor B_T$

is the sketch of $S \cup T$
K-Min Value (KMV) Synopsis

- Given set $S$ of values. Want number of distinct elements $n := D(S)$ (notation)
- Hashing outputs values uniformly in $[0,1]$

- KMV synopsis is ordered set of $k$ smallest values

\[ L = \{U_{(1)}, U_{(2)}, \ldots, U_{(k)}\} \]

- Unbiased Estimator: $\hat{n}_k^{UB} = (k - 1) / U_{(k)}$
  - Exact error analysis based on theory of order statistics
  - Asymptotically optimal as $k$ becomes large

Distributed Data Management, SoSe 2013, S. Michel

Slide based on PPT slides from Beyer et al. ‘07
(Multiset) Union of Partitions

Take union of values and consider again the k smallest ones:

- Combine KMV synopses: $L_A \oplus L_B$
- Theorem: $L$ is a KMV synopsis of $A \cup B$
(Multiset) Intersection of Partitions

- $L = LA \oplus LB$ as before (union): contains $k$ elements
  - $L$ corresponds to a uniform random sample of DVs in $A \cup B$

- $K_{\cap} = \#$ values in $L$ that are also in $D(A \cap B)$

- $K_{\cap} / k$ estimates Jaccard coefficient:
  $$\hat{n}_{\cup} = \frac{|D(A \cap B)|}{|D(A \cup B)|}$$

- Unbiased estimator of #DV\s in the intersection:
  $$\hat{n}_{\cap} = \frac{K_{\cap}}{k} \left( \frac{k-1}{U_{(k)}} \right)$$

$D(set) = $ distinct values

Slide based on PPT slides from Beyer et al. ’07
Min-Hashing

• Hash function $h$ maps elements of a set to integer space. Let’s do that for two sets, $A$ and $B$

• Let $h_{\text{min}}(A)$ and $h_{\text{min}}(B)$ denote the minimum of these numbers for set $A$ and $B$, respectively.

• Then

\[
P[h_{\text{min}}(A) = h_{\text{min}}(B)] = \frac{|A \cap B|}{|A \cup B|}
\]

Jaccard Coefficient
Min-Hashing (Cont’d)

- **Why does it work?** If the min values are the same, that element that causes the min value has to be in \( A \cap B \); probability is \( \frac{|A \cap B|}{|A \cup B|} \)

Unbiased estimator (but too high variance)

- **As seen before for other estimators:** improved estimation quality through multiple “rounds” of estimates (Error is \( O(1/\sqrt{k}) \), with \( k \) rounds)
  - one min value, multiple hash functions
  - several min values, one hash function
Min-Hash: Estimator

count = 0

for each hash function h:
    if \( h_{\text{min}}(A) = h_{\text{min}}(B) \) then
        count++

end

Estimate of Jaccard is given as \( \frac{\text{count}}{k} \)

See exercise in assignment sheet 5
Literature

• Kevin S. Beyer, Peter J. Haas, Berthold Reinwald, Yannis Sismanis, Rainer Gemulla: On synopses for distinct-value estimation under multiset operations. SIGMOD Conference 2007: 199-210
DATA STREAM MANAGEMENT SYSTEMS AND CQL

Thanks to Johannes Gehrke (Cornell) for providing some the following material. Many of the slides are initially based on material by Jennifer Widom (Stanford).
Data Stream Model

• A stream \( S \) is a (possibly) infinite bag (multiset) of elements \(<s, \tau>\) where \( s \) is a tuple belonging to the schema of \( S \) and \( \tau \) is the timestamp of the element.

• Think: tuples of a relational DBMS extended with timestamp, streaming in.
Data Streams: Example

• Monitoring of highway traffic:
  \[ \text{PosSpeedStr(vehicleId, speed, xPos, dir, hwy)} \]

• E.g., for:
  – congestion prediction/warning
  – estimates of travel time
  – toll collection!
  – ticket for too fast driving
Data Streams: Example

• Environmental Monitoring

  StationStream(humidity, solarRadiation, windSpeed, snowHeight)

• Various application scenarios:
  – avalanche risk level computation
  – insights for agriculture
  – air pollution (urban) monitoring
Continuous Queries

• In contrast to ad-hoc, single time queries in (relational) DBMS.

• Queries over Streams are considered continuous: registered once, run “forever”:
  – “want to stay updated to avalanche risk, not just check once”

• Also called standing queries or subscriptions (in publish/subscribe context)

• For instance:
  – Compute average temperature.
  – Select all orders of stock “Apple” with quantity larger than 100.
What and How can we Compute DB-Style Queries?

• How to compute average values over an infinite stream? Block forever?
• How to join infinite streams if join partners can arbitrarily arrive (or not)?

• Idea: keep window that renders a continuous (infinite) stream a snapshot/static relation
Sliding Window Concept

- Focus attention to latest values of stream
- Allows computation of aggregates
- Joins are computed across windows overlaid of other (or same) streams
Sliding Window: Example

• Window of size $W$
  – based on time (=> time-based)
  – or number of tuples inside (count-based)

• Shifted every $t$ by $B$
Sliding Window Aggregates

• Output average for each window when it slides.

• Here:
  – 17.7°C
  – 26.3°C
  – 19.1°C

18.3°C
13.5°C
27.0°C
11.6°C
29.6°C
39.7°C
24.2°C
11.5°C
12.7°C
27.9°C
....
Sliding Window Joins

- Join is executed over individual window contents.
Types of Sliding Windows

• **Time based Window**
  – window contains tuples within a certain time range; e.g., Twitter Tweets of the last 10 minutes, stock market values of the last 10 seconds
  – size can arbitrarily change if input rate changes

• **Count-based Window**
  – window contains at any time a fixed amount of items, say, the last 100 Tweets or 10000 last stock trades
  – newly arriving items kick out older ones (once window is filled up), depending on strategy (next slide)
Types of Sliding Windows (Cont’d)

• **Sliding Window**: move window on certain ticks/time, continuous or in blocks

• **Tumbling Window**: create new window for each time range or size W (i.e., collect data until full or for W time; then reset)

• At each slide/”tumple” a function can be applied to window content and the result outputted

• This is also called “trigger”.

Distributed Data Management, SoSe 2013, S. Michel
Overview of Data Stream Management Systems (DSMSs)

- **STREAM** (Stanford University), **Aurora** (Brandeis/Brown/MIT), **TelegraphCQ** (UC Berkely), **Cayuga** (Cornell), **PIVES** (Uni Marburg), ...

- Large interest also from companies/startups: Oracle, Microsoft, IBM, Streambase

- Lately open-source product for big data distributed streams: Yahoo! S4, Twitter Storm (will see in detail later)
StreamBase Example UI

http://www.streambase.com
STREAM

• Stanford Stream Data Manager
• “General purpose” DSMS for streams and stored data

• Declarative query language to phrase continuous queries (SQL like).

Continuous Query Language – **CQL**

SQL with:

- Streams
- Windows
- New semantics (stream)
  - Three relation-to-stream operators: Istream, Dstream, Rstream
- Sampling

Slide based on material from Jennifer Widom.
Example Relation (Used Later)

**Simplified Linear Road Setup:**

• A single input stream: The stream of positions and speeds of vehicles

  \[ PosSpeedStr(\text{vehicleId}, \text{speed}, \text{xPos}, \text{dir}, \text{hwy}) \]

• vehicleId: vehicle
• speed: speed in MPH
• xPos: Position of the vehicle within the highway in feet
• dir: direction (east or west)
• hwy: highway number

Slide based on material from Jennifer Widom.
Example Query 1

• Two streams:
  – Orders (orderID, customer, cost)
  – Fulfillments (orderID, clerk)

• Total cost of orders fulfilled over the last day by clerk “Sue” for customer “Joe”

\[
\text{SELECT } \text{sum(O.cost)} \\
\text{FROM Orders O, Fulfillments F [Range 1 Day]} \\
\text{WHERE O.orderID = F.orderID and F.clerk = “Sue”} \\
\text{and O.customer = “Joe”}
\]
Example Query 2

- Using a 10% sample of the fulfillments stream, take the 5 most recent fulfillments for each clerk and return the maximum cost.

```
SELECT F.clerk, max(O.cost)
FROM orders O,
    fulfillments F [PARTITION BY clerk ROW 5] 10% SAMPLE
WHERE O.orderID = F.orderID
GROUP BY F.clerk
```
CQL: Relations and Streams

- **T**: discrete, ordered time domain

- A **relation** **R** is a mapping from time **T** to bag of tuples belonging to the schema of **R**.
- **That is**, **R(t)** varies over time
- Updates carry timestamps, too!

- A **stream** is a set of **(tuple, timestamp)** elements
Streams $\leftrightarrow$ Relations

Window specification

Special operators:
Istream, Dstream, Rstream

Any relational query language

Slide based on material from Jennifer Widom.
Stream $\rightarrow$ Relation

- $S [W]$ is a relation - at time $T$ it contains all tuples in window $W$ applied to stream $S$, up to time $T$.

- When $W = \infty$, it contains all tuples in stream $S$ up to time $T$.

- Ways to construct these windows “$[W]$”
  - Time-based
  - Tuple-based
  - Partitioned

Slide based on material from Jennifer Widom.
Time-Based Window

- $S$ [Range $T$]
  - $S$ [Now]
  - $S$ [Range Unbounded]

Examples:
- `PosSpeedStr [RANGE 30 Seconds]`
- `PosSpeedStr [NOW]`
- `PosSpeedStr [RANGE Unbounded]`

Note: variable number of records in the window

Slide based on material from Jennifer Widom.
Tuple-Based Window

- $S \left[ \text{Rows } N \right]$
  - If tuples form a partial order, ties are broken arbitrarily
  - $\left[ \text{Rows Unbounded} \right]$

Example:
- $\text{PosSpeedStr} \left[ \text{ROWS 1} \right]$
Partitioned Windows

- $S \ [\text{Partition By } A_1, \ldots, A_k \ \text{Rows N}]$
  1. Logically partition $S$ into substreams (compare to SQL GROUP By)
  2. Compute a tuple sliding window
  3. Take union

Example: Recall: PosSpeedStr$(vehicleId, speed, xPos, dir, hwy)$
- PosSpeedStr$[\text{PARTITION BY } vehicleId \ \text{ROWS 1}]$

Slide based on material from Jennifer Widom.
Relation ↪ Relation

• With previous window transform we get a relation, now we can apply
• any query expressed in SQL
  – just that deal now with time-varying relations

Example:

• `SELECT distinct vehicleId
  FROM PosSpeedStr [RANGE 30 Seconds]`

*Computes the active vehicles*
Relation $\Rightarrow$ Stream

- $\text{Istream}(R)$ contains a stream element $(r,t)$ whenever $r$ in $R(t) \setminus R(t-1)$ \textbf{“Insert stream”}
- $\text{Dstream}(R)$ contains a stream element $(r,t)$ whenever $r$ in $R(t-1) \setminus R(t)$ \textbf{“Delete stream”}
- $\text{Rstream}(R)$ contains a stream element $(r,t)$ whenever $r$ in $R(t)$ \textbf{“Relation stream”}
Istream, Dstream, and Rstream

- **Istream(R):** contains all tuples in R that are new within the last time period, i.e., insert stream
- **Dstream(R):** contains all tuples in R which were in the stream before the last period (and not anymore in now), i.e., delete stream
- **Rstream(R):** contains all tuples in R

*Note: Istream and Dstream are expressible with Rstream and suitable selections. How?*
Relation $\rightarrow$ Stream: Examples

```
SELECT Istream(*)
FROM PosSpeedStr [RANGE Unbounded]
WHERE speed > 65

SELECT Rstream(*)
FROM PosSpeedStr [NOW]
WHERE speed > 65
```
Query Results at Time T

- Use all relations at time T
- Use all streams up to T, converted to relations
- Compute relational results
- Convert result to streams if desired
Examples

• What is the following query doing?

SELECT Istream(Avg(A)) FROM S [Range 5 seconds]

*Intention maybe: Emit 5-second moving average on every timestep, but output is generated only if average changes (Istream!)*

• To emit a result on every timestep

SELECT Rstream(Avg(A)) FROM S [Range 5 seconds]

• To emit a result on every second

SELECT Rstream(Avg(A)) FROM S
[Range 5 seconds Slide 1 second]

Slide based on material from Jennifer Widom.
Examples (Cont’d)

```
SELECT F.clerk, max(O.cost)
FROM O [∞], F [Rows 1000]
WHERE O.orderID = F.orderID
GROUP BY F.clerk
```

- At time T: entire stream O and last 1000 tuples of F as relations
- Evaluate query, update result relation at T

Orders (orderID, customer, cost)
Fulfillments (orderID, clerk)

Slide based on material from Jennifer Widom.
Examples (Cont’d)

\[
\text{SELECT} \; \text{Istream}(F.\text{clerk}, \max(O.\text{cost})) \\
\text{FROM} \; O \; [\infty], \; F \; [\text{Rows 1000}] \\
\text{WHERE} \; O.\text{orderID} = F.\text{orderID} \\
\text{GROUP BY} \; F.\text{clerk}
\]

- At time T: entire stream O and last 1000 tuples of F as relations
- Evaluate query, update result relation at T
- \textbf{Streamed result:} New result \(<\text{clerk, max}>\), T), whenever \(<\text{clerk, max}>\) changes from T-1

Orders (orderID, customer, cost) 
Fulfillments (orderID, clerk)
Query Execution in STREAM

• When a continuous query is registered, generate a query execution plan
  – New plan merged with existing plans
  – Users can also create & manipulate plans directly

• Plans composed of three main components:
  – Operators
  – Queues (input and inter-operator)
  – State (windows, operators requiring history)

• Global scheduler for plan execution
Simple Query Plan

Slide courtesy of Jennifer Widom.
More Topics

• Seen only formal model and standard concepts of data stream management systems
• There is of course much more to it
• Implementation, optimization (e.g., equivalences), load shedding, ...
• Will be an own lecture by itself.
• Next, look at system aspects in distributed data stream management systems and (mobile) sensor networks
Literature

• http://infolab.stanford.edu/~widom/cql-talk.pdf
• Jürgen Krämer, Bernhard Seeger: Semantics and implementation of continuous sliding window queries over data streams. ACM Trans. Database Syst. 34(1) (2009)