Optimizarea interogărilor în procesarea fluxurilor de date

Query optimization in data stream processing

Rezumat

Sabina Surdu
Conducător științific: Prof. univ. Dr. Leon Țâmbulea

Facultatea de Matematică și Informatică
Universitatea Babeș-Bolyai

Cluj-Napoca

2012

Teza conține următoarele capitole[1]:

[1]Nu detaliem secțiunile capitolelor în acest rezumat.

Cuvinte cheie: fluxuri de date, interogări continue, sisteme de gestiune a fluxurilor de date, optimizarea interogărilor, reducerea consumului de resurse, optimizarea performanței, aplicații pervazive, calcul pervaziv, gestiunea datelor eterogene

Publicații conexate cu teza de doctorat

Rezultatele cercetării și contribuțiile originale prezentate în teză au fost publicate în jurnale sau volume de proceedings ale conferințelor internaționale la care am participat (una dintre lucrări este în curs de apariție):

- **Sabina Surdu** și Vasile-Marian Scuturici, Addressing resource usage in stream processing systems: sizing window effect, IDEAS'11 Proceedings - 15th International Database Engineering & Applications Symposium, paginile 247-248, Lisabona, 2011. Simpozionul este indexat în categoria B în cea mai recentă ierarhizare a conferințelor realizată de Excellence in Research for Australia (ERA), în 2010 [Era10]. (URL articol: `http://dl.acm.org/citation.cfm?id=2076623.2076658&coll=DL&dl=ACM&CFID=63572418&CFTOKEN=57655636`)

- Yann Gripay, Frédérique Laforest, François Lesueur, Nicolas Lumineau, Jean-Marc Petit, Vasile-Marian Scuturici, Samir Sebahi și **Sabina Surdu**, ColisTrack: Testbed for a Pervasive Environment Management System, EDBT 2012 - The 15th International Conference on Extending Database Technology, Berlin, 2012. Conferința e clasificată A de ERA în 2010 [Era10]. (URL lucrări acceptate: `http://edbticdt2012.dima.tu-berlin.de/program/EDBT-papers/`)

- **Sabina Surdu**, A new framework for evaluating performance in data stream monitoring applications with StreamInsight: StreamEval, MaCS 2012 - Booklet of abstracts from The 9th Joint Conference on Mathematics and Computer Science (conferință internațională), pagina 92, Siófok, 2012. (URL Booklet of abstracts: `http://macs.elte.hu/downloads/abstracts/booklet.pdf`)

- **Sabina Surdu**, A New Architecture Supporting The Sizing Window Effect With StreamInsight, Studia Universitatis Babeș-Bolyai Series Informatica, LVI(4):111-120, 2011. Revista este cotată B+ (indexată BDI) de CNCSIS în 2011 [CNC11].

- **Sabina Surdu**, Data stream management systems: a response to large scale scientific data requirements, Annals of the University of Craiova, Mathematics and Computer Science Series, 38(3):66-75, 2011. Revista este cotată B+ (indexată BDI) de CNCSIS în 2011 [CNC11].

- **Sabina Surdu**, A new architecture for load shedding on data streams with StreamInsight: StreamShedder, University of Pitești Scientific Bulletin, Series Electronics and Computers Science, 11(2):57-64, 2011. Revista este cotată B+ (indexată BDI) de CNCSIS în 2011 [CNC11].

- **Sabina Surdu**, A technique for constructing training sets in data stream mining: kSiEved Window Training Set, MDIS 2011 - Proceedings of The Second International Conference on Modelling and Development of Intelligent Systems, paginile 180-191, Sibiu, 2011. (URL volum conferință: `http://conferences.ulbsibiu.ro/mdis/2011/Doc/Proceeding_mdis2011.pdf`)

- **Sabina Surdu**, Towards an education monitoring platform based on data stream processing, Education and Creativity for a Knowledge Society International Conference, The fifth edition - Computer Science Section, paginile 61-66, București, 2011. (URL program conferință: `http://www.utm.ro/conferinta_2011/files/program_conferinta_2011.pdf`)

- **Sabina Surdu**, Online political communication, Interdisciplinary New Media Studies Conference Proceedings (conferință internațională), paginile 55-58, Cluj-Napoca, 2009. (URL program conferință: `http://journalism.polito.ubbcluj.ro/inms/wp-content/uploads/2010/07/INMS_conference_prog.pdf`)

Următoarele manuscrise sunt în curs de evaluare sau urmează a fi trimise la conferințe sau jurnale:

- **Sabina Surdu**, Yann Gripay, Jean-Marc Petit și Vasile-Marian Scuturici, Material trimis la o conferință internațională A* 2012, în curs de evaluare.
- **Sabina Surdu**, A new framework for evaluating performance in data stream monitoring applications with StreamInsight: StreamEval, Annales Universitatis Scientiarum Budapestinensis de Rolando Eötvös Nominatae - Sectio Computatorica, 2012, în curs de evaluare. Lucrarea extinsă a fost trimisă împreună cu abstractul cu același titlu, acceptat la o conferință internațională menționată anterior.
- **Sabina Surdu** și Vasile-Marian Scuturici, Assessing performance in data stream processing, material în lucru pentru IDEAS 2012 - The 16th International Database Engineering & Applications Symposium, Praga, 2012. Simpozionul e clasificat B de ERA în 2010 [Era10].
- **Sabina Surdu**, Data stream processing: traditional vs. dedicated systems (SQL Server vs. StreamInsight), material în lucru pentru Studia Universitatis Babeș-Bolyai Series Informatica. Revista este cotată B+ (indexată BDI) de CNCSIS în 2011 [CNC11].

1 Structura tezei

În Capitolul 1 descriem succint problematica generală a procesării fluxurilor de date cu ajutorul interogărilor continue și realizăm o scurtă incursiune în societatea rețea ubicuă, caracterizată de procesarea datelor eterogene în cadrul mediilor și aplicațiilor pervazive. Prezentăm problema generală a optimizării interogărilor pe fluxurile de date din mediile pervazive și sintetizăm cele două direcții de cercetare pe care le tratăm în această teză: optimizarea consumului de resurse în procesarea interogărilor pe fluxurile de date și gestiunea datelor eterogene în dezvoltarea aplicațiilor pervazive. Aceste aplicații integrează date statice, fluxuri și funcționalități [GLP10]. Prezentăm contribuțiile originale din această teză și menționăm lucrările publicate în jurnale sau prezentate la conferințe internaționale și publicate în volume de proceedings. Enumerăm lucrările pe care le-am trimis la conferințe sau jurnale și care sunt în curs de evaluare sau în curs de apariție.

În Capitolul 2 prezentăm domeniul procesării fluxurilor de date. Introducem sisteme de procesare a fluxurilor de date de referință și discutăm abordări alternative în optimizarea interogărilor, orientate cu precădere către reducerea consumului de resurse ale sistemului, oferind totodată și o viziune comparativă asupra acestora.

În Capitolul 3 descriem tehnicile de optimizare a consumului de resurse în procesarea fluxurilor de date, pe care le propunem în această teză. Analizăm efectul dimensionării ferestrei, în scopul determinării unei dimensiuni de fereastră optime pentru o interogare, astfel încât nivelul resurselor consumate să rămână cât mai redus, iar cerințele de acuratețe să fie respectate. Discutăm tehnica kSiEved Window Training Set, o strategie pentru construirea seturilor de training pentru procesele de *data mining* pe fluxurile de date, ce urmărește de asemenea să reducă utilizarea resurselor în condițiile îndeplinirii cerințelor de acuratețe.

În Capitolul 4 discutăm arhitecturile *resource-aware* pe care le-am proiectat în vederea reducerii consumului de resurse în procesarea interogărilor continue. StreamShedder și WindowSized sunt două astfel de arhitecturi pentru SGFD-uri, bazate pe un sistem comercial de procesare a fluxurilor. Descriem pe scurt StreamEval,

o aplicație ce evaluează variațiile de performanță când condițiile din mediu se schimbă. Discutăm succint SCIPE și InstantSchoolKnow, două propuneri pentru Sisteme de Gestiune a Fluxurilor de Date care vizează domenii de aplicație particulare.

În Capitolul 5 avansăm către dezvoltarea aplicațiilor pervazive. Prezentăm *testbed*-ul pe care l-am realizat în echipă, la LIRIS, INSA Lyon, pentru un sistem care gestionează mediile pervazive, bazat pe un scenariu proiectat pentru astfel de medii, într-un context medical. Acest *testbed* poate fi utilizat pentru analiza dezvoltării aplicațiilor pervazive. Prezentăm designul unei aplicații pervazive centrate pe date, utilizând sistemul SoCQ [GFLP09], tratând într-o manieră omogenă datele din mediul pervaziv. Descriem aplicația pe care am realizat-o pentru scrierea interogărilor continue care combină date eterogene.

În Capitolul 6 evaluăm dezvoltarea aplicațiile pervazive, utilizând mai multe sisteme. Descriem *benchmark*-ul propus și realizăm un studiu experimental. Rezultatele cercetării prezentate în acest capitol fac obiectul unei lucrări pe care am trimis-o la o conferință internațională și care este în prezent în curs de evaluare.

În Capitolul 7 sintetizăm rezultatele obținute pe cele două direcții de cercetare distincte: optimizarea consumului de resurse în procesarea interogărilor pe fluxurile de date și gestiunea datelor eterogene în dezvoltarea aplicațiilor pervazive. Ne oprim asupra tehnicilor de optimizare a consumului de resurse în contextul procesării fluxurilor de date și a arhitecturilor *resource-aware* pe care le-am realizat pentru economisirea resurselor în procesarea interogărilor continue pe fluxuri de date. Discutăm *testbed*-ul pentru dezvoltarea aplicațiilor pervazive, precum și *benchmark*-ul definit pentru evaluarea acestor aplicații. Descriem direcții viitoare de cercetare prilejuite de rezultatele obținute.

2 Procesarea fluxurilor de date în contextul mediilor pervazive

În ultimii ani am asistat la evoluția paradigmei tradiționale de procesare a datelor, de la modelul consacrat, în care datele au o natură statică, la un model dinamic, care cuprinde date caracterizate de o dinamicitate apreciabilă. Într-un număr crescând de domenii, informația se prezintă sub forma fluxurilor continue de date. Acestea reprezintă secvențe potențial infinite de date, care nu pot fi gestionate eficient de SGBD-urile clasice [ACC+03]. O serie de prototipuri pentru administrarea și procesarea fluxurilor de date au fost realizate de echipe din mediul academic. Acestea poartă denumirea de Sisteme de Gestiune a Fluxurilor de Date[2] (SGFD). Industria contribuie la rândul ei la dezvoltarea acestui domeniu, prin proiectarea și dezvoltarea SGFD-urilor (un exemplu recent în acest sens este StreamInsight, realizat de Microsoft [KDA+10]).

Datele din bazele de date tradiționale au o natură statică. Sunt stocate sub forma unor seturi de date finite, care sunt interogate atunci când este necesar [ABB+04]. Pe de altă parte, fluxurile de date sunt dinamice prin însăși definiția lor. Nu sunt stocate permanent în sistem. O interogare în acest context se execută continuu, pe date temporare, care intră în sistem, sunt procesate și în final eliminate. Un SGFD poate executa un număr considerabil de interogări continue complexe [ABB+03], ce iau în calcul mai multe fluxuri de date. Frecvența cu care datele ajung pe flux poate varia în timp. Resursele limitate ale sistemului trebuie să facă față acestor cerințe, în contextul în care procesarea datelor trebuie să ia în calcul și dimensiunea lor temporală.

În aplicațiile din societatea rețea ubicuă[3] [Mur09] individul interacționează nu doar cu alți utilizatori, ci și cu obiecte din mediu, echipate cu dispozitive computaționale [Uni05]. În acest context, fluxurile coexistă cu date modelate în

[2]Termenul consacrat în literatura de specialitate, în limba engleză, este *Data Stream Management System*.

[3]Termenul consacrat în literatura de specialitate, în limba engleză, este *ubiquitous network society*.

maniere diferite. Sistemele de gestiune a acestor medii trebuie să considere, pe lângă fluxuri, și date statice sau funcționalități; mediile pervazive sunt constituite din astfel de elemente și sunt utilizate pentru a modela cât mai fidel realitatea care ne înconjoară [GLL+12]. O integrare a capacităților de interogare a datelor statice, a fluxurilor de date și a funcționalităților într-un cadru unitar, declarativ, deschide alte perspective în procesul de optimizare a interogărilor, în acest nou context, dar similare cu cele din bazele de date tradiționale, bazate pe limbaje asemănătoare cu SQL [Gri09].

Fluxurile de date și aplicațiile pervazive dezvoltate în contextul mediilor pervazive sunt noile componente ale scenariilor din societatea rețea ubicuă. Consumul eficient al resurselor în procesarea fluxurilor de date și dezvoltarea ușoară a aplicațiilor pervazive sunt condiții necesare pentru punerea în practică a societății rețea ubicue.

În acest rezumat redăm graficele, diagramele de sistem sau capturile de ecran așa cum le-am publicat în lucrări de specialitate, în limba engleză. Descriem succint cele mai semnificative contribuții originale prezentate în teza de doctorat.

3 Identificarea problemei

În această teză investigăm problema generală a optimizării interogărilor, în contextul procesării fluxurilor de date continue din mediile pervazive. Identificăm două direcții de cercetare principale, concretizate în publicațiile amintite în preambulul acestui rezumat:

- optimizarea consumului de resurse în procesarea interogărilor pe fluxurile de date;
- investigarea gestiunii datelor eterogene în dezvoltarea aplicațiilor pervazive.

4 Optimizarea consumului de resurse în procesarea interogărilor pe fluxurile de date

Una dintre problemele stringente cu care se confruntă designerii de sisteme pentru procesarea fluxurilor de date este consumul intensiv de resurse ale sistemului. Un sistem care deține resurse limitate trebuie să poată gestiona un număr semnificativ de surse de date, volume de date considerabile, frecvențe impresionante ale fluxurilor, precum și variații imprevizibile ale ritmului în care datele ajung la sistem, așa cum evidențiem în [SS11]. Atât numărul de surse de date, cât și frecvențele fluxurilor, respectiv volumele de date, sunt într-o continuă creștere. Distribuția datelor poate fi variabilă, iar sistemul trebuie să poată executa mai multe interogări complexe [ABB+03], într-o manieră continuă. În acest context, ridicăm problema funcționării corespunzătoare a sistemului în aceste circumstanțe și a evaluării performanței sistemului.

În teza de doctorat prezentăm soluțiile inovatoare pe care le-am propus și care răspund acestei probleme, orientate pe două direcții de cercetare: (1) tehnici de optimizare a consumului de resurse în procesarea fluxurilor de date și (2) dezvoltarea arhitecturilor *resource-aware* pentru procesarea fluxurilor de date, orientate către reducerea consumului de resurse ale sistemului.

Enumerăm în continuare contribuțiile originale pe care le aducem în teza de doctorat, orientate către reducerea consumului de resurse în procesarea fluxurilor de date.

4.1 Efectul dimensionării ferestrei

Dezvoltăm efectul dimensionării ferestrei, *the sizing window effect*, o abordare ce urmărește să optimizeze consumul de resurse la nivelul memoriei și al procesorului, prin calcularea unei dimensiuni de fereastră optime pentru o anumită interogare. Dorim să perfecționăm această tehnică, astfel încât calculul dimensiunii optime să poată fi realizat complet automat de către sistem. Nu cunoaștem niciun

alt studiu anterior care să fi luat în considerare dimensiunea ferestrei input pentru reducerea consumului de resurse. Nu tratăm cazul ferestrelor semantice din punct de vedere temporal (de exemplu, o interogare care calculează viteza medie a vehiculelor pe un segment de drum, în ultimele cinci minute, are nevoie de o fereastră *sliding* semantică de dimensiune fixă). Semantica acestor ferestre este derivată din dimensiunea lor temporală. În cazul nostru, semantica ferestrei nu este conexată cu acest parametru.

Prezentăm pe scurt efectul dimensionării ferestrei (în teză formalizăm riguros domeniul temporal, fluxurile de date, noțiunea de echivalență a interogărilor, rezultatul ideal, rezultatul aproximat, funcția distanță și alte concepte utilizate; în acest rezumat le prezentăm sumar). O fereastră *sliding* este în acest context o porțiune contiguă de date de pe un flux S [BBD+02]. Dacă granițele ei temporale sunt momentele t_i și t_j, vom nota această fereastră cu $SW_{ij}(S)$.

Fie Q o interogare a cărei execuție produce în timp un flux de rezultate agregate. $t_c \in T$ este *timestamp*-ul curent, unde T este domeniul temporal ales. $t_i \in T$ este un *timestamp* care marchează începutul unei ferestre în timp, iar t_0 este *timestamp*-ul primului element emis pe fluxul S. Inițial $t_i = t_c$. Notăm cu $CrtTS$ mulțimea tuturor valorilor *timestamp* din T, pe care le ia t_c. Parcurgem următoarele etape:

1. Stabilim o limită de acuratețe ϵ. Pentru a obține interogări echivalente și a avea răspunsuri valide, diferența între rezultatele ideale și cele aproximate nu trebuie să depășească limita de acuratețe.

2. Calculăm rezultatul ideal R_{s_c} al interogării Q executate pe fluxul de date S, la momentul curent t_c:

$$R_{s_c} = Q(S, t_c) = Q(SW_{0c}(S), t_c), R_{s_c} \in \mathbb{R} \tag{1.1}$$

   unde $SW_{0c}(S)$ este o fereastră *sliding* pe fluxul de date S, ale cărei granițe temporale sunt t_0 și t_c. Numim acest rezultat *ideal*, întrucât ia în considerare toate elementele sosite pe flux până la momentul temporal curent.

3. Descrește constant t_i. Calculăm rezultatele aproximate $R_{w_{c_{\sigma_{ic}}}}$ ale interogării Q executate pe ferestrele *sliding* $SW_{ic}(S)$, la momentul curent t_c. Pentru fiecare valoare temporală t_i, dimensiunea ferestrei $SW_{ic}(S)$ este σ_{ic}, reprezentând numărul de momente temporale conținute în fereastră:

$$R_{w_{c_{\sigma_{ic}}}} = Q(SW_{ic}(S), t_c), R_{w_{c_{\sigma_{ic}}}} \in \mathbb{R} \tag{1.2}$$

Calculăm distanțele între rezultatul ideal și rezultatul aproximat, cu ajutorul unei funcții distanță, pentru fiecare valoare a *timestamp*-ului t_i:

$$distance_{agg_{c_{\sigma_{ic}}}}(R_{s_c}, R_{w_{c_{\sigma_{ic}}}}) = \mid R_{s_c} - R_{w_{c_{\sigma_{ic}}}} \mid \tag{1.3}$$

4. Repetăm pașii 2 și 3 pentru toate valorile *timestamp*-ului curent t_c din $CrtTS$.

5. După finalizarea pasului 4 (când lui t_c i-au fost atribuite toate valorile din $CrtTS$), calculăm media distanțelor între rezultatele aproximate și rezultatele ideale în timp, pentru fiecare dimensiune de fereastră σ_{ic}:

$$AvgDistance(\sigma_{ic}) = \frac{\sum\limits_{t_c \in CrtTS} distance_{agg_{c_{\sigma_{ic}}}}(R_{s_c}, R_{w_{c_{\sigma_{ic}}}})}{|CrtTS|} \tag{1.4}$$

Dimensiunea de fereastră optimă pentru interogarea Q este cea mai mică dimensiune de fereastră pentru care distanța medie între rezultatele aproximate și rezultatele ideale este sub limita de acuratețe ϵ.

În experimentele realizate pe un set de interogări agregate utilizăm datele din *benchmark*-ul Linear Road [ACG+04]. Sunt date simulate, referitoare la traficul rutier pe drumuri expres. Fiecare drum este divizat în 100 de segmente.

Pentru fiecare dintre următoarele interogări agregate vom aplica tactica descrisă anterior.

Interogarea 1: Calculează numărul mediu de vehicule pe unitatea de timp care au călătorit pe un anumit segment până în prezent.

Interogarea 2: Calculează viteza medie pe un anumit segment.

Interogarea 3: Calculează taxa medie plătită de un vehicul (pe toate segmentele).

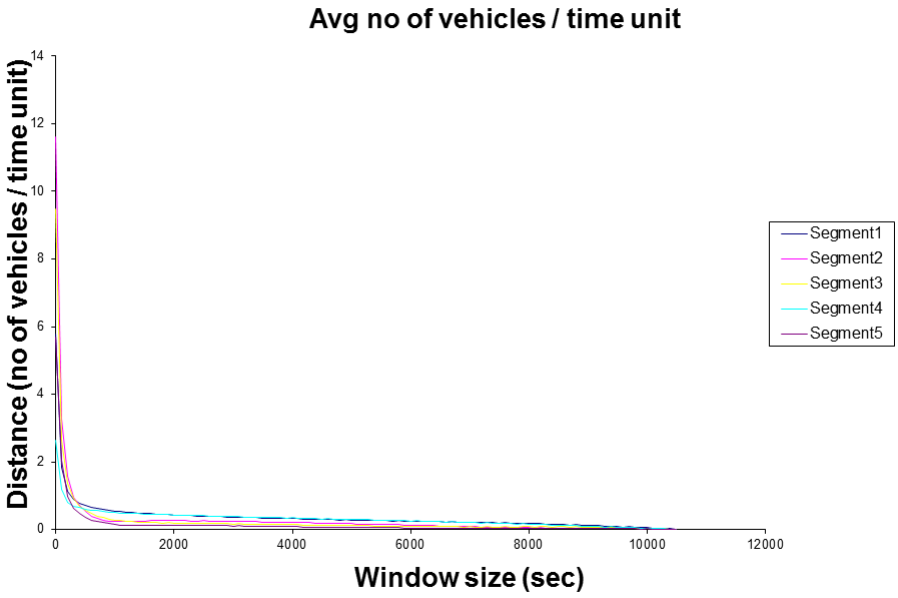


Figura 1.1: Numărul mediu de vehicule pe unitatea de timp, calculat separat pentru cinci segmente.

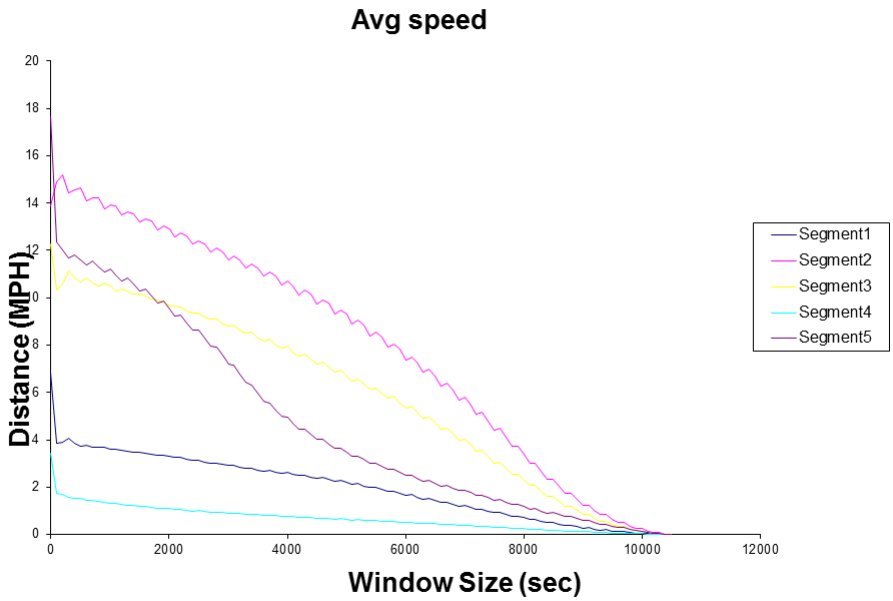


Figura 1.2: Viteza medie, calculată separat pentru cinci segmente

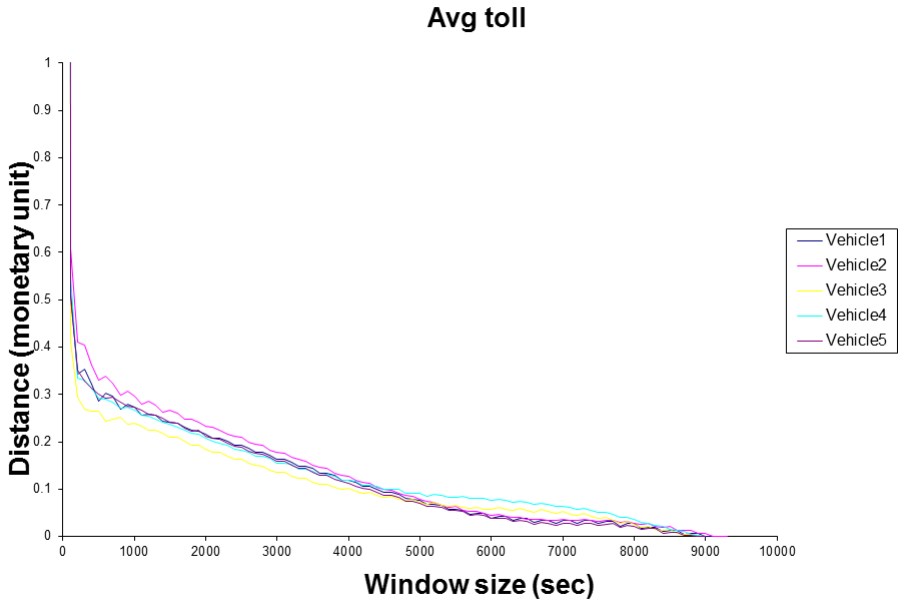


Figura 1.3: Taxa medie plătită de 5 vehicule (calculată separat pentru fiecare vehicul).

În cazul Interogării 1 (figura 1.1, pe care am publicat-o în [Sur11a]) observăm că distanța medie între rezultatele ideale și rezultatele aproximate este sub limita de acuratețe 1 pentru orice dimensiune de fereastră care depășește 1000 de momente temporale.

În cazul Interogării 2 (figura 1.2) distanța medie între rezultatele ideale și rezultatele aproximate este sub limita de acuratețe 1 pentru orice dimensiune de fereastră care depășește 10000 de momente temporale.

În cazul Interogării 3 (figura 1.3) distanța medie între rezultatele ideal și rezultatele aproximate este sub limita de acuratețe 0.1 pentru orice dimensiune de fereastră care depășește 6000 de momente temporale.

Administratorul unei aplicații care implementează Linear Road poate specifica o limită de acuratețe pentru Interogarea 1 pe datele output (rezultate ale interogării calculate pe ferestre *sliding*), astfel încât acestea să nu difere cu mai mult de 1 față de rezultatul ideal. Sistemul poate să execute această interogare pe o fereastră de 1000

de momente temporale. Constrângeri similare pot fi formulate și pentru celelalte interogări. Rezultatele acestei cercetări sunt publicate în [SS11].

4.2 kSiEved Window Training Set

Una dintre provocările întâlnite în procesul de *data mining* este aplicarea tehnicilor de *data mining* pe fluxuri de date continue [ZB03]. Dezvoltăm o tehnică ce ia în considerare resursele sistemului în construirea seturilor de training pentru algoritmii de *data mining* pe fluxuri de date, și anume tehnica kSiEved Window Training Set (kSEWT), prima metodă care "cerne" un flux de date în funcție de anumiți parametri, pentru a construi seturi de training în acest context, respectând cerințele de acuratețe. Definim un nou model de date, modelul kSiEved, care se bazează pe ferestre kSiEved, construite din ferestre *sliding* prin aplicarea unor funcții de extragere a pozițiilor dintr-o fereastră, definite riguros în teză.

kSEWT calculează rezultate corecte, pe ferestre *sliding* SW_{ic}, la fiecare moment temporal t_c (omitem fluxul S în definiția acestor ferestre pentru a simplifica notațiile). Pentru fiecare astfel de fereastră, kSEWT construiește ferestre kSiEved $SEW_{ic}(k)$, pe baza unui parametru k, care variază în timp. Acesta din urmă generează o "sită" cu orificii care va "cerne" elementele ferestrei SW_{ic}, realizând fereastra kSiEved $SEW_{ic}(k)$, pe care se calculează de asemenea rezultate ale interogărilor. kSEWT estimează acuratețea rezultatelor obținute pe ferestre kSiEved față de rezultatul corect utilizând o funcție distanță. În funcție de media distanțelor calculate, este ales parametrul k (valoarea maximă a acestuia), pentru care media distanțelor față de rezultatul corect nu depășește o limită admisă a erorii δ. Parametrul k furnizează Setul de Training kSiEved Window (kSiEved Window Training Set), constituit din mulțimea tuturor ferestrelor kSiEved de parametru k obținute în experiment.

Prezentăm rezultatele experimentale obținute pe un set de date cu o distribuție uniformă. Aplicând kSEWT am obținut graficul din figura 1.4. Dacă alegem o limită $\delta = 0.5$, din acest grafic reiese că putem aplica "site" cu parametrul k = 2, când con-

struim setul de training. Aceasta înseamnă că renunțăm la jumătate dintre tuplurile input, în condițiile respectării cerințelor de acuratețe enunțate, ceea ce înseamnă o economie substanțială de resurse. Această cercetare este publicată în [Sur11b].

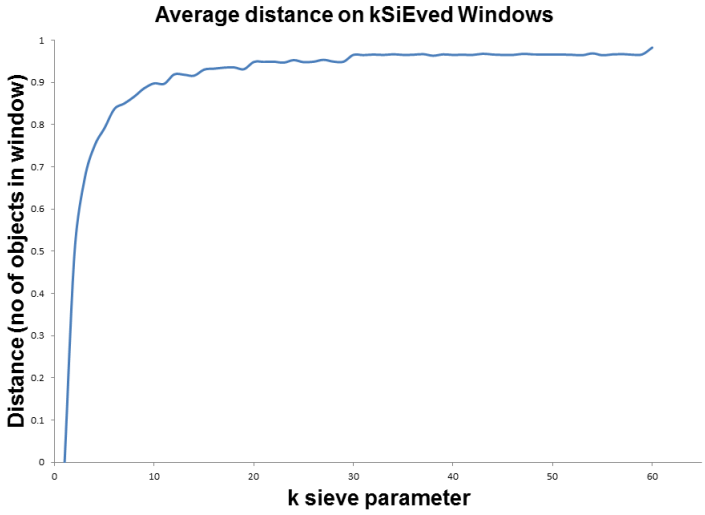


Figura 1.4: Distanța medie între rezultatele corecte și rezultatele interogării pe ferestrele kSiEved

4.3 WindowSized

În continuarea demersului de cercetare centrat pe efectul dimensionării ferestrei, propunem o nouă arhitectură *resource-aware* pentru implementarea acestui efect, utilizând Microsoft StreamInsight [KDA+10]: WindowSized. Principala contribuție a acestei arhitecturi este reprezentată de integrarea modulului WindowSizing într-o aplicație de monitorizare dezvoltată cu StreamInsight. WindowSizing interacționează cu motorul de interogări, cu interfețele către sursele de date - pentru a modifica dimensiunea ferestrei și cu interfețele către dispozitivele output - pentru a obține rezultatele interogărilor.

Figura 1.5 înfățișează principalele componente ale unei astfel de arhitecturi. Nivelul inferior al arhitecturii (cu *Event sources*, *Input adapters*, *StreamInsight query engine*, *Output adapters* și *Event targets*) este preluat din arhitectura propusă de Mi-

crosoft pentru implementarea unei aplicații cu StreamInsight [SIA]. WindowSized e bazată pe principiile de proiectare ale unei aplicații tipice StreamInsight, cu următoarele elemente: surse de date, adaptori input, interogări continue pe server, adaptori output și consumatori de date [GSK+09]. Contribuția noastră este reprezentată de integrarea modulului WindowSizing într-o astfel de arhitectură. Rezultatele obținute sunt publicate în [Sur11a].

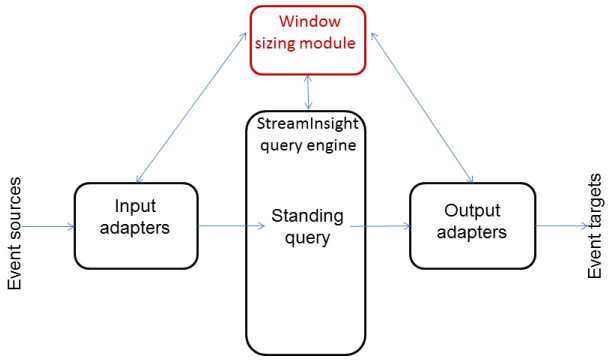


Figura 1.5: Arhitectura WindowSized

4.4 StreamShedder

Dezvoltăm aplicația StreamingTraffic pentru monitorizarea traficului. Propunem o nouă arhitectură pentru o astfel de aplicație de monitorizare realizată utilizând platforma Microsoft StreamInsight [KDA+10]. Dezvoltăm modulul de *load shedding* [ABB+04] StreamShedder și recomandăm integrarea acestuia în arhitectura aplicației StreamingTraffic. StreamShedder realizează operații de eliminare a datelor într-o manieră parametrizată, luând în considerare resursele sistemului și timpul de răspuns al interogărilor. Arhitectura rezultată integrează strategii de *load shedding* cu un sistem comercial de procesare a fluxurilor de date pentru a obține performanțe superioare în procesarea interogărilor continue.

Figura 1.6 înfățișează arhitectura modificată a unei aplicații de monitorizare implementate cu StreamInsight, care cuprinde modulul StreamShedder. La fel ca în cazul WindowSized, nivelul inferior al arhitecturii (cu *Event sources*, *Input adapters*,

StreamInsight query engine, *Output adapters* și *Event targets*) este preluat din arhitectura propusă de Microsoft pentru implementarea unei aplicații cu StreamInsight [SIA]. Contribuția noastră este reprezentată de integrarea modulului StreamShedder într-o astfel de arhitectură. Vom denumi această arhitectură îmbunătățită tot StreamShedder, pe baza modulului care realizează operațiile de *load shedding*.

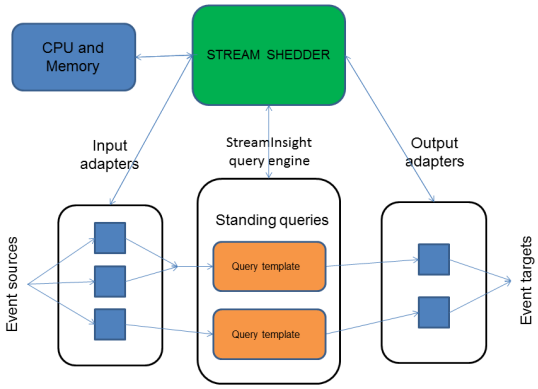


Figura 1.6: Arhitectura StreamShedder

StreamShedder este un modul software implementat în C#. Comunică cu un modul de monitorizare a memoriei și procesorului (CPU and Memory), care furnizează resursele de sistem utilizate. Pe baza limitelor de memorie și timp de procesor specificate de utilizator, StreamShedder poate ordona adaptorilor input să elimine anumite tupluri. StreamShedder monitorizează și timpul de răspuns al interogărilor de pe server. Pe baza datelor pe care le primește, comunică adaptorilor input ce tupluri să elimine. Pentru a evalua impactul asupra semanticii aplicației, acest modul comunică și cu adaptorii output, obținând rezultatele interogărilor. Rezultatele acestei cercetări sunt publicate în [Sur11d].

4.5 StreamEval

Dezvoltăm o soluție care evaluează variațiile de performanță când diferite condiții din mediu se schimbă (de exemplu, când manipulăm frecvența surselor de date care alimentează interogările continue de pe server): StreamEval. În implementarea aplicației de monitorizare și a *framework*-ului propus utilizăm platforma co-

mercială menționată anterior în acest rezumat, dezvoltată de Microsoft în ultimii ani: StreamInsight [AGR+09]. Utilizăm aplicația de monitorizare StreamingTraffic dezvoltată în contextul arhitecturii StreamShedder (secțiunea 4.4).

Notăm cu DR (*data rate*) frecvența fluxului input, definită ca numărul de elemente care ajung pe fluxul input S în fiecare secundă. Vom utiliza această notație când modificăm frecvența sursei de date. Folosim notația (ușor modificată) *ConsumedGate* din [Mon] pentru a ne referi la punctul imediat următor adaptorilor input (la primul operator dintr-o interogare continuă Q).

Utilizăm atributele de monitorizare a interogărilor oferite de API-ul ManagementService [Mon]. Ca în [Mon], suntem interesați în monitorizarea timpului de răspuns mediu consumat (*average consumed latency*), între două momente temporale t_1 and t_2. Prin urmare, evaluăm numărul de tupluri procesate $TupleCount$ și timpul de răspuns $Latency$ la *ConsumedGate*, la momentele t_1 și t_2. Notăm timpul de răspuns mediu consumat cu $AvgLat$ (*average latency*). Calculăm $AvgLat$ aplicând formula din [Mon]:

$$AvgLat = (Latency_{t_2} - Latency_{t_1})/(TupleCount_{t_2} - TupleCount_{t_1}). \quad (1.5)$$

Modificăm frecvența sursei de date după cum urmează. Începem cu valoarea 1 pentru DR (un eveniment pe secundă) și evaluăm valoarea $AvgLat$ corespunzătoare. Mărim DR până la 500 evenimente pe secundă. Măsura $AvgLat$ rămâne sub o milisecundă. Chiar și pentru valori ale DR de 1000 de evenimente / secundă, care depășesc cerințele StreamingTraffic, $AvgLat$ se menține în jurul aceleiași valori. Această cercetare este descrisă sumar în abstractul [Sur12a]. Lucrarea extinsă e în curs de evaluare [Sur12b].

4.6 SCIPE

Dezvoltăm un set de principii, SCIPE (*SCIentific data stream processing PrinciplEs*), orientat către realizarea unui Sci-SGFD, un Sistem de Gestiune a Fluxurilor de Date în contextul datelor de dimensiuni foarte mari din domenii care țin de științele

exacte. Comunitățile de cercetare din științele exacte lucrează cu seturi de date de ordinul petaocteților, iar pentru viitorul apropiat se preconizează dimensiuni de câțiva exaocteți [BLW09]. În acest context investigăm posibilitatea realizării unui SGFD pliat pe necesitățile comunităților din științele exacte. Studierea obiectivelor domeniului cercetat poate conduce la optimizarea consumului de resurse în interogările continue pe fluxurile de date.

Redăm în continuare setul de principii SCIPE:

1. Când situația o permite, se procesează, iar ulterior se sumarizează sau se elimină un element. Acest principiu are un impact semnificativ asupra consumului de memorie, menținând elementele sub forma unui sumar, dacă este necesară procesarea lor ulterioară.

2. Dacă este necesară stocarea individuală a elementelor, se rețin doar acelea din trecutul recent și se elimină sau se sumarizează elementele vechi.

3. Se proiectează un sistem care conține modalități de revizuire a elementelor (strategie reținută din [AAB+05]).

4. Se realizează operațiuni de *load shedding* într-o manieră semantică, dependentă de domeniul de aplicație (un exemplu de sistem care realizează *load shedding* semantic este Aurora [ACC+03]).

5. Se construiesc interogările într-un mod atractiv pentru utilizator, combinând limbaje vizuale și o interfață declarativă SQL (se observă aici îmbinarea abordărilor din [ACC+03] și [ABW06]).

SCIPE și motivarea acestui demers de cercetare sunt publicate în [Sur11c].

4.7 InstantSchoolKnow

Analizăm domeniul educațional și modalitățile în care utilizarea fluxurilor de date poate conduce la optimizarea proceselor educaționale. Realizăm EdStream, un set de reguli care pot fi aplicate în realizarea unei platforme de monitorizare educaționale bazate pe procesarea fluxurilor de date. Propunem designul unei

platforme de monitorizare educaționale, InstantSchoolKnow. Scopul acesteia este să achiziționeze continuu date de la instituții de învățământ (înregistrate în cadrul platformei), să realizeze analiza acestor date utilizând paradigma procesării continue și să publice rezultatele acestei analize în timp real. Pentru a atinge acest obiectiv trebuie parcurse următoarele etape: înregistrarea pe platforma InstantSchoolKnow, achiziția datelor, analiza datelor și publicarea datelor. Spre deosebire de abordările curente, InstantSchoolKnow își propune să unifice funcționalități de *e-learning* și monitorizare a elevilor într-o singură platformă. Această cercetare este publicată în [Sur11e].

4.8 O platformă pentru accesarea datelor de pe dispozitive mobile inteligente

Propunem o arhitectură pentru realizarea unei platforme online cu conținut orientat către comunicarea politică. În faza inițială datele au o natură statică și pot fi accesate de pe dispozitive mobile inteligente. Dorim să extindem această platformă *new media* cu funcții de procesare a fluxurilor de date și serviciilor, în contextul unui mediu pervaziv. Această cercetare este publicată în [Sur09].

5 Gestiunea datelor eterogene în dezvoltarea aplicațiilor pervazive

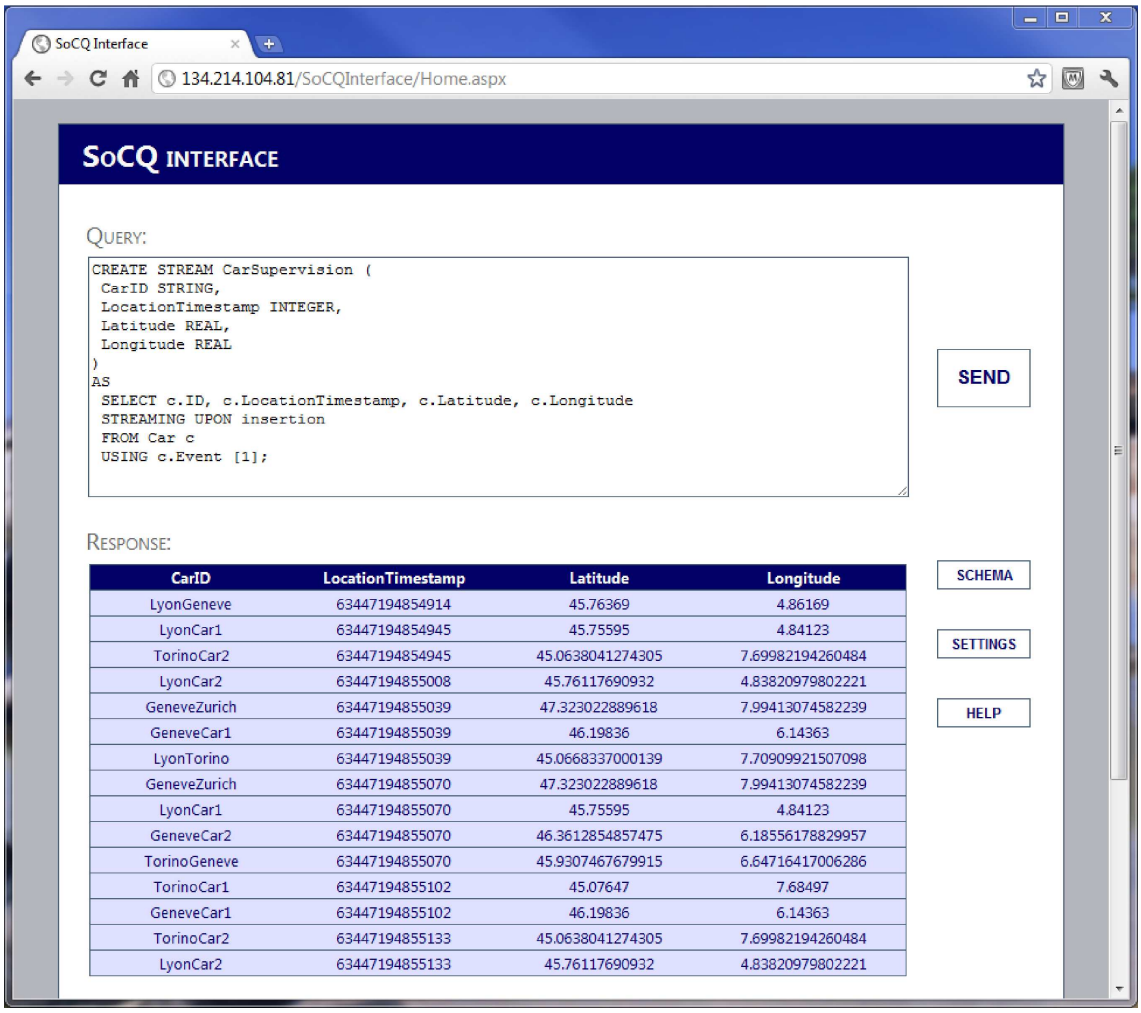
Un număr considerabil de scenarii și de aplicații pervazive bazate pe aceste scenarii sunt constituite din date statice (similare cu cele din bazele de date tradiționale), fluxuri de date și funcționalități sau servicii distribuite [GLP10], în conformitate cu situațiile reale din viața de zi cu zi pe care le modelează. Pentru a gestiona toate aceste elemente dintr-un mediu pervaziv, se recurge de cele mai multe ori la programarea *ad hoc*[4], care integrează mai multe paradigme de programare (limbaje imperative, limbaje declarative și protocoale de rețea) [Gri09]. Soluțiile dezvoltate în această manieră sunt însă dificil de implementat și se realizează în perioade lungi de timp. Investigăm variante alternative pentru implementarea aplicațiilor pervazive și metode de evaluare a procesului de dezvoltare.

Enumerăm în continuare contribuțiile originale pe care le-am adus în contextul gestiunii datelor eterogene în aplicațiile pervazive, în teza de doctorat.

5.1 Gestiunea datelor eterogene într-un mediu pervaziv

Abordăm una dintre principalele provocări din calculul pervaziv: înlesnirea dezvoltării aplicațiilor pervazive. Descriem un scenariu pentru monitorizarea unor containere într-un context medical, ce implică transportul conținutului medical în recipiente echipate cu senzori. Pe baza acestui scenariu, discutăm un *testbed* util în dezvoltarea aplicațiilor și evaluarea procesului de dezvoltare și arătăm cum se poate construi o aplicație pervazivă, utilizând sistemul SoCQ (Service-oriented Continuous Query) [GFLP09]. Scenariul, simularea scenariului ca testbed, vizualizarea sa și aplicația pervazivă realizată reprezintă contribuțiile intrinseci ale acestui demers de cercetare, pe care le-am dezvoltat în cadrul echipei cu care am lucrat la LIRIS, INSA Lyon. Rezultatele cercetării fac obiectul unui articol acceptat la o conferință internațională, aflat în curs de publicare [GLL+12].

[4]Termenul consacrat în literatura de specialitate, în limba engleză, este *ad hoc programming*.



CarID	LocationTimestamp	Latitude	Longitude
LyonGeneve	63447194854914	45.76369	4.86169
LyonCar1	63447194854945	45.75595	4.84123
TorinoCar2	63447194854945	45.0638041274305	7.69982194260484
LyonCar2	63447194855008	45.76117690932	4.83820979802221
GeneveZurich	63447194855039	47.323022889618	7.99413074582239
GeneveCar1	63447194855039	46.19836	6.14363
LyonTorino	63447194855039	45.0668337000139	7.70909921507098
GeneveZurich	63447194855070	47.323022889618	7.99413074582239
LyonCar1	63447194855070	45.75595	4.84123
GeneveCar2	63447194855070	46.3612854857475	6.18556178829957
TorinoGeneve	63447194855070	45.9307467679915	6.64716417006286
TorinoCar1	63447194855102	45.07647	7.68497
GeneveCar1	63447194855102	46.19836	6.14363
TorinoCar2	63447194855133	45.0638041274305	7.69982194260484
LyonCar2	63447194855133	45.76117690932	4.83820979802221

Figura 1.7: Aplicația Web care permite scrierea interogărilor continue

Pentru a interacționa cu motorul de interogări, implementăm o aplicație Web ASP.NET. Aceasta permite unui dezvoltator să scrie interogări continue, ce combină date eterogene din mediu, utilizând un limbaj de interogări asemănător cu SQL, specific sistemului SoCQ [Gri09]. Dacă dorim să monitorizăm în fiecare moment pozițiile fiecărei mașini, scriem o interogare în acest limbaj, care generează ca rezultat toate locațiile mașinilor în timp real. Figura 1.7 înfățișează aplicația Web, o interogare și rezultatele acesteia.

5.2 AgilBench

Propunem un *benchmark* pentru evaluarea dezvoltării aplicațiilor pervazive. Utilizăm mai multe sisteme în acest sens și realizăm un studiu experimental. Rezultatele acestei cercetări au fost incluse într-o lucrare pe care am trimis-o la o conferință internațională și care este în prezent în curs de evaluare [SGPS12].

6 Concluzii și direcții viitoare de cercetare

Cele două direcții de cercetare urmate s-au concretizat, după cum am arătat, în dezvoltarea unor arhitecturi, strategii și tehnici pentru optimizarea consumului de resurse în procesarea fluxurilor de date, dar și în realizarea unui *testbed* și a unui *benchmark* în contextul dezvoltării aplicațiilor pervazive. Aceste contribuții au fost publicate în reviste sau volume de proceedings ale unor conferințe internaționale. Un material este acceptat pentru publicare, iar alte două materiale sunt în curs de evaluare.

Domeniul fluxurilor de date și al aplicațiilor pervazive se află într-o continuă dinamică. În mod previzibil, propunerile noastre vor suferi modificări în timp. Intenționăm să automatizăm efectul dimensionării ferestrei, astfel încât sistemul să poată alege automat dimensiunea optimă a ferestrei și să perfecționăm arhitecturile *resource-aware* propuse, astfel ca toate deciziile să fie luate de sistem, fără intervenția utilizatorului. Dorim să adăugăm noi servicii în *testbed*-ul propus și să îmbogățim *benchmark*-ul pe care l-am realizat pentru evaluarea dezvoltării aplicațiilor pervazive. Evaluăm posibilitatea de a proiecta un sistem capabil să gestioneze mediile pervazive, care să permită înlocuirea totală a scenariului care modelează un mediu pervaziv, fără nicio schimbare în implementare. Cele mai performante sisteme (cum ar fi SoCQ) au nevoie de noi module în cazul în care mecanismele de acces la date se schimbă odată cu înlocuirea scenariului.

Bibliografia tezei

[AAB+05] Daniel J. Abadi, Yanif Ahmad, Magdalena Balazinska, Ugur Cetintemel, Mitch Cherniack, Jeong-Hyon Hwang, Wolfgang Lindner, Anurag S. Maskey, Alexander Rasin, Esther Ryvkina, Nesime Tatbul, Ying Xing și Stan Zdonik. The Design of the Borealis Stream Processing Engine. În *CIDR 2005, Proceedings of Second Biennial Conference on Innovative Data Systems Research*, paginile 277–289, 2005.

[ABB+03] Arvind Arasu, Brian Babcock, Shivnath Babu, Mayur Datar, Keith Ito, Rajeev Motwani, Itaru Nishizawa, Utkarsh Srivastava, Dilys Thomas, Rohit Varma și Jennifer Widom. STREAM: The Stanford Stream Data Manager. *IEEE Data Engineering Bulletin*, 26(1):19–26, 2003.

[ABB+04] Arvind Arasu, Brian Babcock, Shivnath Babu, John Cieslewicz, Mayur Datar, Keith Ito, Rajeev Motwani, Utkarsh Srivastava și Jennifer Widom. STREAM: The Stanford Data Stream Management System. Raport tehnic, Stanford InfoLab, 2004.

[ABC+05] Yanif Ahmad, Bradley Berg, Ugur Cetintemel, Mark Humphrey, Jeong-Hyon Hwang, Anjali Jhingran, Anurag Maskey, Olga Papaemmanouil, Alex Rasin, Nesime Tatbul, Wenjuan Xing, Ying Xing și Stanley B. Zdonik. Distributed operation in the Borealis stream processing engine. În *SIGMOD Conference*, paginile 882–884, 2005.

[ABW06] Arvind Arasu, Shivnath Babu și Jennifer Widom. The CQL continuous query language: Semantic foundations and query execution. *The VLDB Journal*, 15(2):121–142, 2006.

[ACC+03] Daniel J. Abadi, Donald Carney, Ugur Cetintemel, Mitch Cherniack, Christian Convey, Sangdon Lee, Michael Stonebraker, Nesime Tatbul și Stanley B. Zdonik. Aurora: a new model and architecture for data stream management. *The VLDB Journal*, 12(2):120–139, 2003.

[ACG+04] Arvind Arasu, Mitch Cherniack, Eduardo Galvez, David Maier, Anurag S. Maskey, Esther Ryvkina, Michael Stonebreaker și Richard Tibbetts. Linear Road: A Stream Data Management Benchmark. În *VLDB'04, Proceedings of The Thirtieth International Conference on Very Large Data Bases*, paginile 480–491, 2004.

[Adm] Federal Highway Administration. Congestion Pricing: A Primer. `http://www.ops.fhwa.dot.gov/publications/congestionpricing/congestionpricing.pdf`.

[Agg07] Charu C. Aggarwal. An Introduction to Data Streams. În *Data Streams - Models and Algorithms*, paginile 1–8. 2007.

[AGR+09] Mohamed H. Ali, Ciprian Gerea, Balan Sethu Raman, Beysim Sezgin, Tiho Tarnavski, Tomer Verona, Ping Wang, Peter Zabback, Asvin Ananthanarayan, Anton Kirilov, Ming Lu, Alex Raizman, Ramkumar Krishnan, Roman Schindlauer, Torsten Grabs, Sharon Bjeletich, Badrish Chandramouli, Jonathan Goldstein, Sudin Bhat, Ying Li, Vincenzo Di Nicola, Xianfang Wang, David Maier, Stephan Grell, Olivier Nano și Ivo Santos. Microsoft CEP Server and Online Behavioral Targeting. *Proceedings of the VLDB Endowment*, 2(2):1558–1561, august 2009.

[AIS93] Rakesh Agrawal, Tomasz Imielinski și Arun Swami. Mining association rules between sets of items in large databases. În *SIGMOD '93,*

Proceedings of the 1993 ACM SIGMOD international conference on Management of data, paginile 207–216, 1993.

[AMT06] Serge Abiteboul, Ioana Manolescu și Emanuel Taropa. A Framework for Distributed XML Data Management. În *EDBT 2006, Proceedings of The 10th International Conference on Extending Database Technology*, paginile 1049–1058, 2006.

[AW04] Arvind Arasu și Jennifer Widom. A Denotational Semantics for Continuous Queries over Streams and Relations. *SIGMOD Record*, 33(3):6–12, 2004.

[BBC+04] Hari Balakrishnan, Magdalena Balazinska, Donald Carney, Ugur Cetintemel, Mitch Cherniack, Christian Convey, Eduardo F. Galvez, Jon Salz, Michael Stonebraker, Nesime Tatbul, Richard Tibbetts și Stanley B. Zdonik. Retrospective on Aurora. *The VLDB Journal*, 13(4):370–383, 2004.

[BBD+02] Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani și Jennifer Widom. Models and Issues in Data Stream Systems. În *PODS*, paginile 1–16, 2002.

[BBD+04] Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani și Dilys Thomas. Operator scheduling in data stream systems. *The VLDB Journal*, 13(4):333–353, 2004.

[BBDM03] Brian Babcock, Shivnath Babu, Mayur Datar și Rajeev Motwani. Chain: Operator Scheduling for Memory Minimization in Data Stream Systems. În *SIGMOD Conference*, paginile 253–264, 2003.

[BBS04] Magdalena Balazinska, Hari Balakrishnan și Michael Stonebraker. Load management and high availability in the Medusa distributed stream processing system. În *SIGMOD '04, Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, paginile 929–930, 2004.

[BDM04] Brian Babcock, Mayur Datar și Rajeev Motwani. Load Shedding for Aggregation Queries over Data Streams. În *ICDE 2004, Proceedings of the 20th International Conference on Data Engineering*, paginile 350–361, 2004.

[BH07] Don Box și Anders Hejlsberg. LINQ: .NET Language-Integrated Query. `http://msdn.microsoft.com/en-us/library/bb308959.aspx`, 2007.

[BLW09] Jacek Becla, Kian-Tat Lim și Daniel Liwei Wang. Report from the 3rd Workshop on Extremely Large Databases. *Data Science Journal*, 8:MR1–MR16, 2009.

[CCC+02] Don Carney, Ugur Cetintemel, Mitch Cherniack, Christian Convey, Sangdon Lee, Greg Seidman, Michael Stonebraker, Nesime Tatbul și Stan Zdonik. Monitoring Streams - a New Class of Data Management Applications. În *VLDB '02, Proceedings of the 28th International Conference on Very Large Data Bases*, paginile 215–226, 2002.

[CCD+03] Sirish Chandrasekaran, Owen Cooper, Amol Deshpande, Michael J. Franklin, Joseph M. Hellerstein, Wei Hong, Sailesh Krishnamurthy, Sam Madden, Vijayshankar Raman, Fred Reiss și Mehul Shah. TelegraphCQ: Continuous Dataflow Processing for an Uncertain World. În *CIDR 2003, Proceedings of the First Biennial Conference on Innovative Data Systems Research*, 2003.

[CCR+03] Don Carney, Ugur Cetintemel, Alex Rasin, Stan Zdonik, Mitch Cherniack și Michael Stonebraker. Operator Scheduling in a Data Stream Manager. În *VLDB '03, Proceedings of the 29th International Conference on Very Large Data Bases*, paginile 838–849, 2003.

[CDTW00] Jianjun Chen, David J. DeWitt, Feng Tian și Yuan Wang. NiagaraCQ: A Scalable Continuous Query System for Internet Databases. În *Proce-*

edings of ACM SIGMOD International Conference on Management of Data, paginile 379–390, 2000.

[CEP] Complex Event Processing. `http://www.complexevents.com/`.

[CG05] Graham Cormode și Minos N. Garofalakis. Sketching Streams Through the Net: Distributed Approximate Query Tracking. În *VLDB 2005, Proceedings of the 31st International Conference on Very Large Data Bases*, paginile 13–24, 2005.

[Cha] Nicholas Chase. The ultimate mashup – Web services and the semantic Web, Part 1: Use and combine Web services. `http://www.ibm.com/developerworks/xml/tutorials/x-ultimashup1/`.

[CNC11] Consiliul Național al Cercetării Științifice din Învățământul Superior. Situația curentă a revistelor recunoscute CNCSIS. `http://www.cncsis.ro/userfiles/file/CENAPOSS/Bplus_2011.pdf`, 2011.

[Cor08] Computing Research and Education. `http://core.edu.au/cms/images/downloads/conference/Astar.pdf`, 2008.

[CSD11] Alfredo Cuzzocrea, Il-Yeol Song și Karen C. Davis. Analytics over large-scale multidimensional data: the big data revolution! În *DOLAP'11, Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*, paginile 101–104, 2011.

[CVC+10] Víctor Cuevas-Vicenttín, Genoveva Vargas-Solar, Christine Collet, Noha Ibrahim și Christophe Bobineau. Coordinating Services for Accessing and Processing Data in Dynamic Environments. În *OTM'10, Proceedings of the 2010 International Conference on On the move to meaningful internet systems - Volume Part I*, paginile 309–325, 2010.

[dDCK+06] Scott de Deugd, Randy Carroll, Kevin E. Kelly, Bill Millett și Jeffrey Ricker. SODA: Service Oriented Device Architecture. *IEEE Pervasive Computing*, 5(3):94–96, 2006.

[DG08] Jeffrey Dean și Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, ianuarie 2008.

[DGIM02] Mayur Datar, Aristides Gionis, Piotr Indyk și Rajeev Motwani. Maintaining Stream Statistics over Sliding Windows. În *SODA 2002, ACM-SIAM Symposium on Discrete Algorithms*, paginile 635–644, 2002.

[ECPS02] Deborah Estrin, David Culler, Kris Pister și Gaurav Sukhatme. Connecting the Physical World with Pervasive Networks. *IEEE Pervasive Computing*, 1(1):59–69, ianuarie 2002.

[Era10] Excellence in Research for Australia 2010 (Australian Research Council). Ranked Conference List. `http://www.arc.gov.au/era/era_2010/archive/key_docs10.htm`, 2010.

[FHA10] Fatima Farag, Moustafa Hammad și Reda Alhajj. Adaptive query processing in data stream management systems under limited memory resources. În *Proceedings of the 3rd workshop on Ph.D. students in information and knowledge management*, paginile 9–16, 2010.

[FHL+11] Nicolas Ferry, Vincent Hourdin, Stephane Lavirotte, Gaetan Rey, Michel Riveill, și Jean-Yves Tigli. Wcomp, a middleware for ubiquitous computing. În *Ubiquitous Computing*, paginile 151–176, 2011.

[FPSS96] Usama M. Fayyad, Gregory Piatetsky-Shapiro și Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3):37–54, 1996.

[GAE06] Thanaa M. Ghanem, Walid G. Aref și Ahmed K. Elmagarmid. Exploiting predicate-window semantics over data streams. *SIGMOD Record*, 35(1):3–8, 2006.

[Geh09] Johannes Gehrke. Technical perspective - Data stream processing: when you only get one look. *Communications of the ACM*, 52(10):96, 2009.

[GFLP09] Yann Gripay, Frédérique Laforest și Jean-Marc Petit. SoCQ: a Pervasive Environment Management System. În *UbiMob'09, 5èmes Journées Francophones Mobilité et Ubiquité*, paginile 87–90, 2009.

[GLL+12] Yann Gripay, Frédérique Laforest, François Lesueur, Nicolas Lumineau, Jean-Marc Petit, Vasile-Marian Scuturici, Samir Sebahi și Sabina Surdu. ColisTrack: Testbed for a Pervasive Environment Management System. În *EDBT 2012, The 15th International Conference on Extending Database Technology. În curs de apariție*, 2012.

[GLP07] Yann Gripay, Frédérique Laforest și Jean-Marc Petit. Towards Action-Oriented Continuous Queries in Pervasive Systems. În *BDA'07, Bases de Données Avancées 2007*, paginile 1–20, 2007.

[GLP09] Yann Gripay, Frédérique Laforest și Jean-Marc Petit. SoCQ: a Framework for Pervasive Environments. În *ISPAN 2009, 10th International Symposium on Pervasive Systems, Algorithms and Networks*, paginile 154–159, 2009.

[GLP10] Yann Gripay, Frédérique Laforest și Jean-Marc Petit. A Simple (yet Powerful) Algebra for Pervasive Environments. În *EDBT 2010, Proceedings of The 13th International Conference on Extending Database Technology*, paginile 1–12, 2010.

[Gooa] Google Maps API Family. `http://code.google.com/apis/maps/index.html`.

[Goob] The Google Directions API. `http://code.google.com/apis/maps/documentation/directions/`.

[Gri08] Yann Gripay. Service-oriented Continuous Queries for Pervasive Systems. În *EDBT 2008 PhD Workshop (unofficial proceedings)*, paginile 1–7, 2008.

[Gri09] Yann Gripay. *A Declarative Approach for Pervasive Environments: Model and Implementation*. Teză de doctorat, Institut National des Sciences Appliquées de Lyon, 2009.

[GS10] Yann Gripay și Vasile-Marian Scuturici. Managing Distributed Service Environments: a Data-oriented approach. În *UbiMob'10, 6èmes Journées Francophones Mobilité et Ubiquité*, paginile 1–4, 2010.

[GSK+09] Torsten Grabs, Roman Schindlauer, Ramkumar Krishnan, Jonathan Goldstein și Rafael Fernández. Introducing Microsoft StreamInsight. Raport tehnic, Microsoft, 2009.

[GZK05] Mohamed Medhat Gaber, Arkady Zaslavsky și Shonali Krishnaswamy. Mining data streams: A review. *ACM SIGMOD Record*, 34(2):18–26, 2005.

[HL11] Martin Hilbert și Priscila Lopez. The World's Technological Capacity to Store, Communicate and Compute Information. *Science*, 332(6025):60–65, februarie 2011.

[HMS01] David J. Hand, Heikki Mannila și Padhraic Smyth. *Principles of Data Mining*, paginile 1–24. The MIT Press, Cambridge, MA, USA, 2001.

[IGLS06] Jon Espen Ingvaldsen, Jon Atle Gulla, Tarjei Laegreid și Paul Christian Sandal. Financial News Mining: Monitoring Continuous Streams of Text. În *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, paginile 321–324, 2006.

[IM06] Edurne Izkue și Eduardo Magana. Sampling time-dependent parameters in high-speed network monitoring. În *PM2HW2N 2006, Proceedings of the ACM International Workshop on Performance Monitoring, Measurement, and Evaluation of Heterogeneous Wireless and Wired Networks*, paginile 13–17, 2006.

[Int] Ovidiu Vermesan, Mark Harrison, Harald Vogt, Kostas Kalaboukas, Maurizio Tomasella, Karel Wouters, Sergio Gusmeroli

și Stephan Haller. Internet of Things. Strategic Research Roadmap. http://www.grifs-project.eu/data/File/CERP-IoT%20SRA_IoT_v11.pdf.

[JMHA10] Oana Jurca, Sebastian Michel, Alexandre Herrmann și Karl Aberer. Continuous query evaluation over distributed sensor networks. În *ICDE'10, Proceedings of The 26th IEEE International Conference on Data Engineering*, paginile 912–923, 2010.

[KDA+10] Seyed J. Kazemitabar, Ugur Demiryurek, Mohamed H. Ali, Afsin Akdogan și Cyrus Shahabi. Geospatial Stream Query Processing using Microsoft SQL Server StreamInsight. *Proceedings of the VLDB Endowment*, 3(2):1537–1540, septembrie 2010.

[KG10] Ramkumar Krishnan și Jonathan Goldstein. A Hitchhiker's Guide to Microsoft StreamInsight Queries. Raport tehnic, Microsoft, iunie 2010.

[Kog07] Jacob Kogan. *Introduction to Clustering Large and High-Dimensional Data*, paginile 98–99. Cambridge University Press, NY, USA, 2007.

[Lan09] Marc Langheinrich. A survey of RFID privacy approaches. *Personal and Ubiquitous Computing*, 13(6):413–421, august 2009.

[Lin] LINQ documentation. http://msdn.microsoft.com/en-us/library/bb397926.aspx.

[LMT+05] Jin Li, David Maier, Kristin Tufte, Vassilis Papadimos și Peter A. Tucker. Semantics and Evaluation Techniques for Window Aggregates in Data Streams. În *SIGMOD Conference*, paginile 311–322, 2005.

[MCP+02] Alan M. Mainwaring, David E. Culler, Joseph Polastre, Robert Szewczyk și John Anderson. Wireless sensor networks for habitat monitoring. În *Proceedings of the 1st ACM International Workshop on Wireless Sensor Networks and Applications*, paginile 88–97, 2002.

[Mea] Text REtrieval Conference (TREC). Common Evaluation Measures, 2011. `http://trec.nist.gov/pubs/trec19/appendices/measures.pdf`.

[Mei11] Erik Meijer. The World According to LINQ. *Communications of the ACM*, 54(10):45–51, octombrie 2011.

[Mon] StreamInsight documentation. Monitoring the StreamInsight Server and Queries. `http://msdn.microsoft.com/en-us/library/ee391166.aspx`.

[Mur09] Teruyasu Murakami. The Age of Ubiquitous. *Highlighting Japan through articles*, 2(10):8–9, februarie 2009.

[MWA+03] Rajeev Motwani, Jennifer Widom, Arvind Arasu, Brian Babcock, Shivnath Babu, Mayur Datar, Gurmeet Singh Manku, Chris Olston, Justin Rosenstein și Rohit Varma. Query Processing, Resource Management, and Approximation in a Data Stream Management System. În *CIDR 2003, Proceedings of the First Biennial Conference on Innovative Data Systems Research*, 2003.

[Nas09] Hebah H. O. Nasereddin. Stream Data Mining. *International Journal of Web Applications*, 1(4):183–190, decembrie 2009.

[Pug08] William Pugh. Technical perspective: A methodology for evaluating computer system performance. *Communications of the ACM*, 51(8):82–82, august 2008.

[RMCZ06] Esther Ryvkina, Anurag S. Maskey, Mitch Cherniack și Stan Zdonik. Revision Processing in a Stream Processing Engine: A High-Level Design. În *ICDE 2006, Proceedings of the 22nd International Conference on Data Engineering*, paginile 141–143, 2006.

[Rys11] Michael Rys. Scalable SQL. *Communications of the ACM*, 54(6):48–53, iunie 2011.

[Sch07] Sven Schmidt. *Quality-of-Service-Aware Data Stream Processing*. Teză de doctorat, Dresden University of Technology, Department of Computer Science, 2007.

[Sch09] Arnd Schröter. Modeling and optimizing content-based publish/subscribe systems. În *Proceedings of the 6th Middleware Doctoral Symposium*, paginile 5:1–5:6, 2009.

[Scu09] Marian Scuturici. Dataspace API. Raport tehnic, LIRIS, septembrie 2009.

[SGPS12] Sabina Surdu, Yann Gripay, Jean-Marc Petit și Vasile-Marian Scuturici. Lucrare în curs de evaluare. Conferință internațională A*, 2012.

[SIA] StreamInsight Server Architecture. `http://msdn.microsoft.com/en-us/library/ee391536.aspx`.

[Sima] Mark Simms. 101'ish LINQ Samples for StreamInsight (part 1 - filtering and aggregation). `http://blogs.msdn.com/b/masimms/archive/2010/09/16/101-ish-linq-samples-for-streaminsight.aspx`.

[Simb] Mark Simms. Using SQL Server for reference data in a StreamInsight query. `http://windowsazurecat.com/2011/08/sql-server-reference-data-streaminsight-query`.

[SM03] Debashis Saha și Amitava Mukherjee. Pervasive Computing: A Paradigm for the 21st Century. *IEEE Computer*, 36(3):25–31, martie 2003.

[Soc] Proiectul SoCQ. `http://socq.liris.cnrs.fr/`.

[SS11] Sabina Surdu și Vasile-Marian Scuturici. Addressing resource usage in stream processing systems: sizing window effect. În *IDEAS'11 Proceedings, 15th International Database Engineering & Applications Symposium*, paginile 247–248, 2011.

[Stra] StreamInsight documentation. Creating Input and Output Adapters. http://msdn.microsoft.com/en-us/library/ee378877.aspx.

[Strb] StreamInsight documentation. Microsoft StreamInsight. http://msdn.microsoft.com/en-us/library/ee362541.aspx.

[Sur09] Sabina Surdu. Online Political Communication. În *Interdisciplinary New Media Studies Conference Proceedings*, paginile 55–58, 2009.

[Sur11a] Sabina Surdu. A New Architecture Supporting The Sizing Window Effect With StreamInsight. *Studia Universitatis Babeș-Bolyai Series Informatica*, LVI(4):111–120, 2011.

[Sur11b] Sabina Surdu. A technique for constructing training sets in data stream mining: kSiEved Window Training Set. În *MDIS 2011, Proceedings of The Second International Conference on Modelling and Development of Intelligent Systems*, paginile 180–191, 2011.

[Sur11c] Sabina Surdu. Data stream management systems: a response to large scale scientific data requirements. *Annals of the University of Craiova, Mathematics and Computer Science Series*, 38(3):66–75, 2011.

[Sur11d] Sabina Surdu. A new architecture for load shedding on data streams with StreamInsight: StreamShedder. *University of Pitești Scientific Bulletin, Series Electronics and Computers Science*, 11(2):57–64, 2011.

[Sur11e] Sabina Surdu. Towards an education monitoring platform based on data stream processing. În *Education and Creativity for a Knowledge Society International Conference, The fifth edition - Computer Science Section*, paginile 61–66, 2011.

[Sur12a] Sabina Surdu. A new framework for evaluating performance in data stream monitoring applications with StreamInsight: StreamEval. În *MaCS 2012, Booklet of abstracts from The 9th Joint Conference on Mathematics and Computer Science*, pagina 92, 2012.

[Sur12b] Sabina Surdu. A new framework for evaluating performance in data stream monitoring applications with StreamInsight: StreamEval. În curs de evaluare la Annales Universitatis Scientiarum Budapestinensis de Rolando Eötvös Nominatae, Sectio Computatorica, 2012.

[SW04] Utkarsh Srivastava și Jennifer Widom. Flexible Time Management in Data Stream Systems. În *PODS '04*, paginile 263–274, 2004.

[TAC+06] Nesime Tatbul, Yanif Ahmad, Ugur Cetintemel, Jeong-Hyon Hwang, Ying Xing și Stanley B. Zdonik. Load Management and High Availability in the Borealis Distributed Stream Processing Engine. În *GSN*, paginile 66–85, 2006.

[Tam03] Leon Țâmbulea. *Baze de date*. Universitatea Babeș-Bolyai, Cluj-Napoca, România, ediția a 6-a, 2003.

[Tat02] Nesime Tatbul. Qos-driven load shedding on data streams. În *EDBT '02, Proceedings of the Workshops XMLDM, MDDE, and YRWS on XML-Based Data Management and Multimedia Engineering-Revised Papers*, paginile 566–576, 2002.

[TCZ+03] Nesime Tatbul, Ugur Cetintemel, Stan Zdonik, Mitch Cherniack și Michael Stonebraker. Load shedding in a data stream manager. În *VLDB '03, Proceedings of the 29th International Conference on Very Large Data Bases*, paginile 309–320, 2003.

[TCZa+03] Nesime Tatbul, Ugur Cetintemel, Stan Zdonik, Mitch Cherniack și Michael Stonebraker. Load Shedding on Data Streams. În *MPDS'03, ACM Workshop on Management and Processing of Data Streams*, 2003.

[Tib03] Richard S. Tibbetts. Linear Road: Benchmarking Stream-Based Data Management Systems. Teză de masterat. Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2003.

[Tpc] TPC Benchmarks. `http://www.tpc.org/information/benchmarks.asp`.

[TTPM02] Pete Tucker, Kristin Tufte, Vassilis Papadimos și David Maier. NEXMark - a benchmark for queries over data streams. Raport tehnic. OGI School of Science and Engineering at OHSU, 2002.

[TZ06] Nesime Tatbul și Stan Zdonik. Window-aware load shedding for aggregation queries over data streams. În *VLDB '06, Proceedings of The 32nd International Conference on Very Large Data Bases*, paginile 799–810, 2006.

[Uni05] International Telecommunication Union. *The Internet of Things*. ITU Internet Reports. International Telecommunication Union, 2005.

[Wei91] Mark Weiser. The Computer for the 21st Century. *Scientific American*, 265(3):94–104, septembrie 1991.

[XL05] Wenwei Xue și Qiong Luo. Action-Oriented Query Processing for Pervasive Computing. În *CIDR 2005, Proceedings of The Second Biennial Conference on Innovative Data Systems Research*, paginile 305–316, 2005.

[XLD] XLDB - Extremely Large Databases. `http://www.xldb.org/`.

[YK96] Qi Yang și Haris N. Koutsopoulos. A microscopic traffic simulator for evaluation of dynamic traffic management systems. *Transportation Research Part C*, 4(3):113–129, 1996.

[ZB03] Qiankun Zhao și Sourav S. Bhowmick. Sequential Pattern Mining: A Survey. Raport tehnic, Nanyang Technological University, Singapore, 2003.

[ZSC+03] Stanley B. Zdonik, Michael Stonebraker, Mitch Cherniack, Ugur Cetintemel, Magdalena Balazinska și Hari Balakrishnan. The Aurora and Medusa Projects. *IEEE Data Engineering Bulletin*, 26(1):3–10, 2003.