

SPADE: The System S Declarative Stream Processing Engine



B.Gedik, H. Andrade, K. Wu, P. Yu, and M. Doo
(SIGMOD. 2008)

Presented by Kenneth Lui (wckl2)
10th Nov 2015

Outline

- Background - Stream Processing Engine, System S
- Motivation
- System Design & Contribution - Programming Model, Optimization
- Example & Experiment Result
- Future Work
- Summary & Critical Analysis

Background



Stream Processing Engine

- “On-the-fly” processing of time ordered series of events or values
 - Low-Latency is key
- Data enter the system as “input stream”, get filtered, processed, aggregated etc. in the network of “computational elements” connected by streams
- Related Works
 - MillWheel (Google), Apache Storm (Twitter)

Stream Processing Use Cases

- Web log processing
- Sensor networks
- Real-time financial analysis

System S

- Large-scale, distributed data stream processing middleware and application development framework
- Applications organized as data-flow graphs
 - Sets of **Processing Elements (PEs)** connected by streams
 - PEs are distributed over the computing nodes
 - Each stream carries a series of **Stream Data Objects (SDOs)**
 - The PE ports and streams connecting them are **typed**
- Provide reliability, scheduling, placement optimization, security, fault tolerance etc.

Stream Processing Core (System S)

- Dataflow Graph Manager (DGM)
 - Define stream connections among PEs
- Data Fabric (DR)
 - Distributed data transport daemons
- Resource Manager (RM)
 - Makes global resource decisions for PEs and streams
- PE Execution Container (PEC)
 - Provide run-time context and security barrier

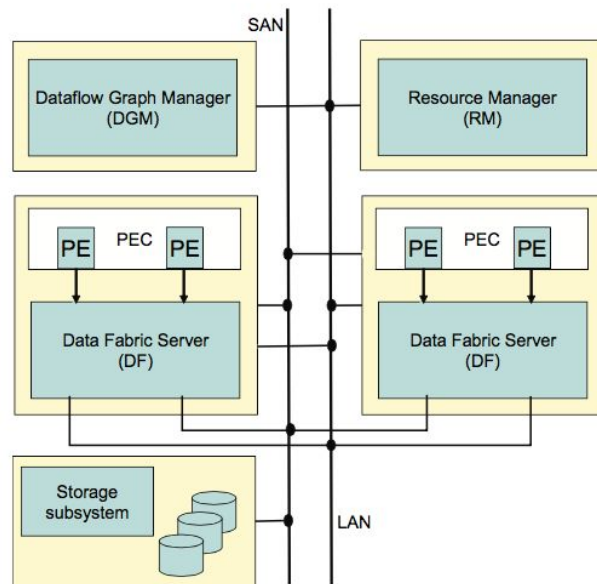


Figure 1: Key components of System S that provide services to run stream applications [18].

Motivation



Before SPADE, there were two ways of use System S...

Programming in PE API

- For experienced developer
- Write programs in **C++** or **Java** to interact directly with PEs
- Design configuration files to specify the topology of the data-flow graph (i.e. connect the PEs)

Working with Domain Specific Queries

- For less experienced developers
- Issue natural language-like domain-specific inquiries
- **Inquiry Services (INQ) planner** makes use of a repository of existing PEs to automatically create a data-flow graph

SPADE - Declarative middle-ground

- SPADE = Stream Processing Application Declarative Engine
- Declarative = Developers describe the problem rather than the steps to solve it
- Allow integration of User defined functions (UDFs) and Legacy Code
- Some manual tuning on deployment is possible

Users: Little/No Expertise

Users: Knowledgeable in Declarative Querying

Users: Experts in Programming



High level Inquiry
E.g.: "Estimate the customer satisfaction"

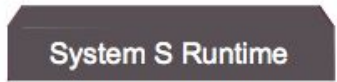
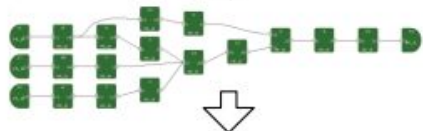


Describe the application using a toolkit of stream processing operators



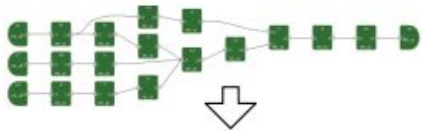
Program PEs
Connect them

```
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000
```



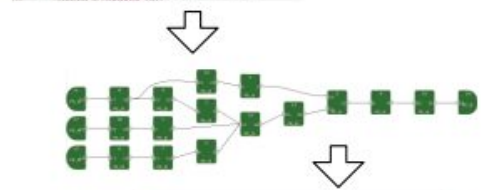
User interacts with an Intelligent System

INQ



User interacts with a High-level Language

SPADE



User interacts with a Programming API

PE API

Figure 1: System S from an application developer's perspective

System Design & Contribution



Code Generation Framework

- Compiler takes query specification written in SPADE's intermediate language and produces these native parts in System S:
 - PE template
 - Node pools
 - PE topology
 - PE binaries
 - Job description (from System S Job Description Language Compiler)

Code Generation Framework

- SPADE compiler's output is highly customized based on the system characteristics
 - Underlying network topology
 - Computer architecture

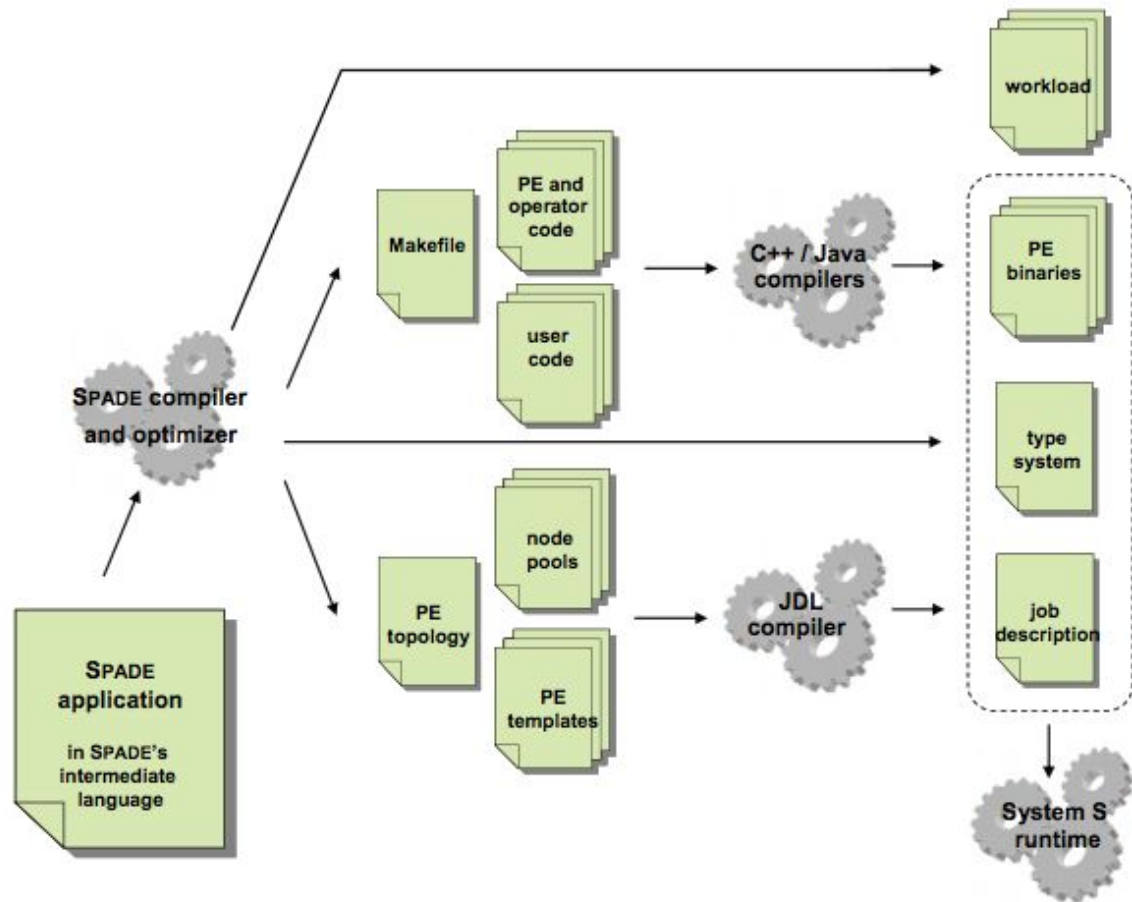


Figure 3: Spade's code generation framework

Stream Processing Operators

- Functor
- Aggregate
- Join
- Sort
- Barrier - used as a synchronization point
- Puncctor - generate punctuation for windowing
- Split
- Delay

Edge Adapters

- Source
 - Parsing
 - Tuple creation
- Sink
 - From streams to external data
 - E.g. file system, network

SPADE Programming Language

```
# %1 and %2 are the first and second parameters
#define NCNT min(%1,16) /* number of nodes to utilize */
#define FCNT min(%2,30) /* number of days to analyze */
```

Application meta-
information

```
[Application]
vwap # trace
```

Type definitions

```
[Typedefs]
namespace vwap
```

Node pools

```
[Nodepools]
nodepool ComputingPool[16] := () # automatically allocated from available nodes
```

```
[Program]
/* Source data format:
 * 1 ticker:String, 8 volume:Float, 15 askprice:Float, 22 peratio:Float,
 * 2 ... */
```

Program body

SPADE Programming Language

```
for_begin @day 1 to FCNT # for each day

    stream TradeQuote@day(ticker:String, ttype:String, price:Float, volume:Float,
askprice:Float, asksize:Float)
        := Source() ["file:///gpfss/taq"+select(@day<10,"0@day","@day")+ ".csv",
nodelays, csvformat] { 1, 5, 7-8, 15-16 }
        -> partition["mypartition_@day"], ComputingPool[mod(@day-1,NCNT)]

    stream TradeFilter@day(ticker: String, myvwap:Float, volume:Float)
        := Functor(TradeQuote@day) [ttype="Trade" & volume>0.0]
        { myvwap := price*volume }
        -> partitionFor(TradeQuote@day), ComputingPool[mod(@day-1,NCNT)]

    stream VWAPAggregator@day(ticker:String, svwap:Float, svolume:Float)
        := Aggregate(TradeFilter@day ) [ticker]
        { Any(ticker), Sum(myvwap), Sum(volume) }
        -> partitionFor(TradeQuote@day), ComputingPool[mod(@day-1,NCNT)]
```

SPADE Programming Language

```
stream BargainIndex@day(ticker:String, bargainindex:Float)
  := Join(VWAP@day ; QuoteFilter@day )
    [{ticker}={ticker}, cvwap > askprice*100.0]
    { bargainindex := exp(cvwap-askprice*100.0)*asksize }
  -> partitionFor(TradeQuote@day), ComputingPool[mod(@day-1,NCNT)]
```

```
export stream NonZeroBargainIndex@day(schemaof(BargainIndex@day))
  := Functor(BargainIndex@day) [bargainindex>0.0] {}
  -> partitionFor(TradeQuote@day), ComputingPool[mod(@day-1,NCNT)]
```

```
Null := Sink(NonZeroBargainIndex@day) ["file:///Bargains@day.dat"]{}
  -> partitionOf(TradeQuote@day), ComputingPool[mod(@day-1,NCNT)]
```

```
for_end
```

User-Defined Operators

- Can make use of external libraries to implement domain-customized operations
- Allow converting legacy code to System S
- Support interfacing with external platforms

Advanced Features

- List Types and Vectorized Operations
- Flexible Windowing Schemes
 - Tumbling windows - fixed number of tuples
 - Sliding windows - expiration policy + trigger mechanism
 - Punctuation-based window boundaries
- Pergroup Aggregates and Joins

Compiler Optimizations

- Operator Grouping
- Execution Model
- Vectorized Processing

Operator Grouping

- Having multiple operators per PE is more efficient
- Reduce message transmission and queuing delays

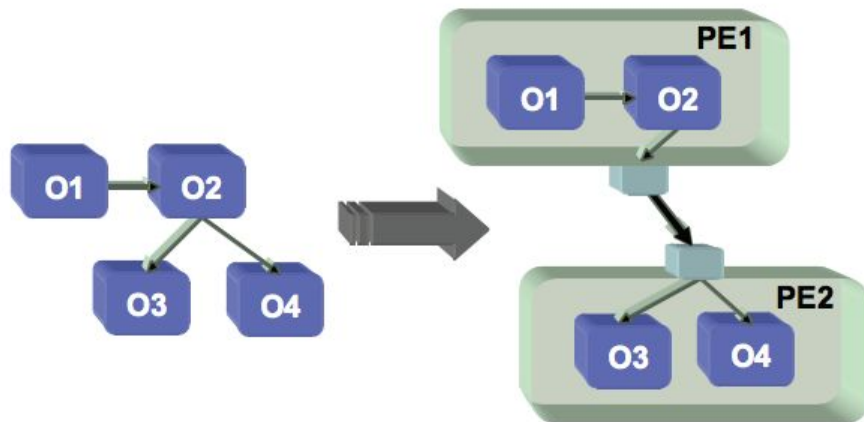


Figure 4: Example operator to PE mapping

Execution Model

- To make use of multiple cores, SPADE create multiple PE's to be run on the same node
- Multi-threading built-in operators were still under development

Vectorized Processing

- Single-Instruction Multiple-Data (SIMD)
- E.g. Intel's Streaming SIMD Extensions (SSE)

Operator Fusion

- Operators in the same PE are chained as depth-first function calls, without any queuing
- For thread-safe operators, SPADE supports multi-threading to cut short the main PE thread
 - May require locking

Two-phase learning-based Optimization

- First, compile the application in a special “Statistics Collection mode”
 - Application is run in this mode to collect metrics like CPU load and network traffic
- Then, compile the application for a second time
 - Optimizer uses statistics to guide operator grouping & fusion to come up with the PEs

Example & Experiment result

Bargain Index Computation

- Compute the bargain index (a scalar metric for stock trading analysis) for every stock symbol that appears in the source stream
- Source: Live stock data can be read directly from the IBM WebSphere Front Office (WFO)
- Sink: IBM DB2 Data Stream Edition – an extension of DB2 designed for persisting high-rate data streams

Bargain Index Computation

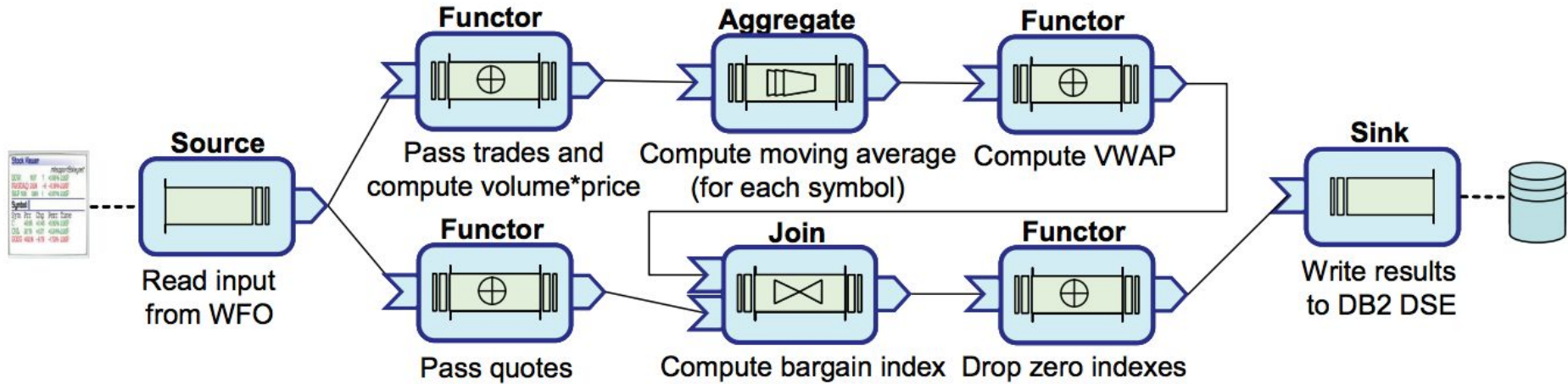


Figure 5: Bargain Index computation for *all* stock symbols

Experiment

- Process 22 days' worth of ticker data for ≈ 3000 stocks with a total of ≈ 250 million trade and quote transactions
- ≈ 20 GBs of data, sharded per file per day on the disk on IBM's General Parallel File System (GPFS)
- Parallelize the processing by running 22 instances (PEs), one for each trading day, over 16 nodes in our cluster

Issues with this experiment

- All operators within the same query are packed into a single PE (i.e. single PE per day)
- No inter-node communication or cooperation
- Some resources are idle after ~23:07
- Compare with native System S API implementation?

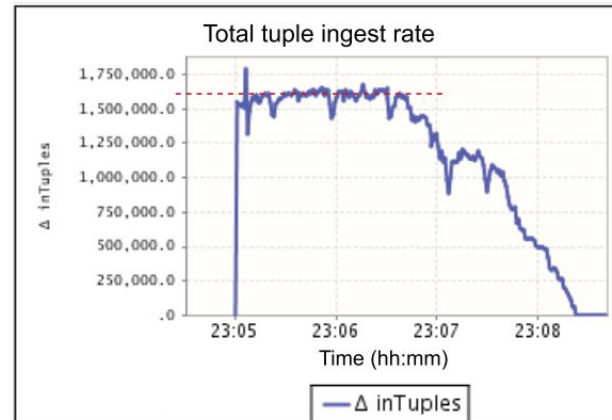


Figure 6: Tuple ingestion rate for the parallel and distributed bargain index computation application, using 22 parallel queries distributed over 16 nodes.

Future Work



Future Work

- Visual development environment
- Domain-specific operator
 - (e.g. signal processing, stream data mining)
- Higher-level languages (Stream SQL, semantic composition framework)
 - A 2013 paper about “IBM Streams Processing Language (SPL)”
- Interoperability
 - Data ingestion and externalization with other platforms

Summary & Critical Analysis



Summary

- A declarative language which balances flexibility and barrier of entry
- Toolkit (compiler, stream operators)
- Bring stream processing to System S

Critical Analysis - System

- Partition and optimization happen at compile-time
- Does not adopt to capacity change (+/- nodes)
- No priority concept for the tuples

Critical Analysis - Paper

- Two-phase learning-based optimization is not discussed in depth
 - I am very curious about the development/deployment workflow here
 - It should compare the performance with/without this optimization
- No fault tolerance analysis
- Example & Evaluation not representative

Thank you!



Any questions?