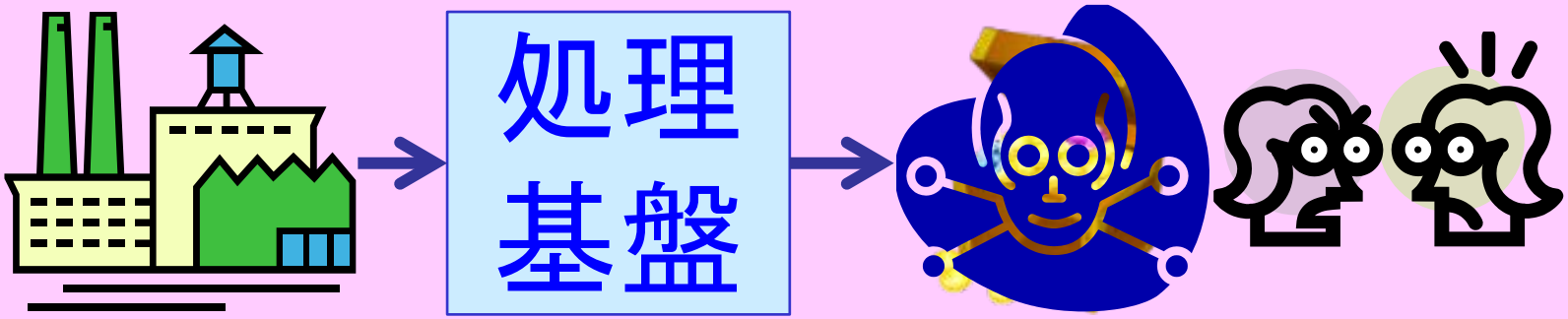
データ管理基盤の
マルウェア分析への適用

筑波大学 システム情報系・計算科学研究センター
川島英之

Big Data

- 制御・観測系センサ
  - **マルウェア検知**，自動車，工場
- 通信ログ
  - CRS-3(322Tbps)，通信会社(数千万人），広告
- 監視カメラ
  - Ring of Steel（ロンドン，NY）
- E-Science
  - 衛星画像(170TB)，粒子衝突器(15PB/Year)，望遠鏡(20TB/night)

データ処理基盤で 攻撃検知



データ処理パラダイム

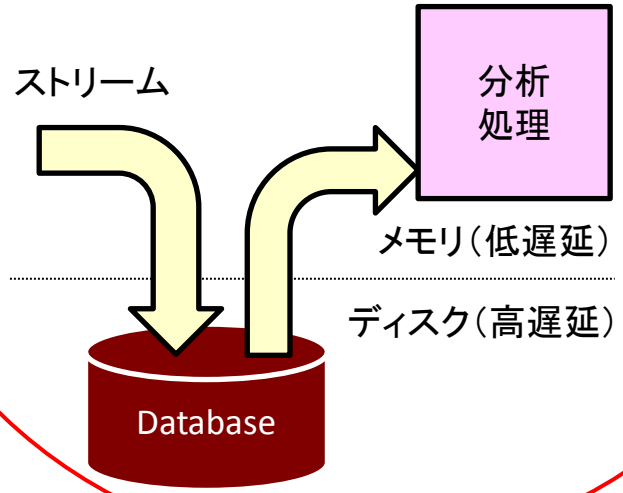
バッチ（DBMS）/ リアルタイム（DSMS）

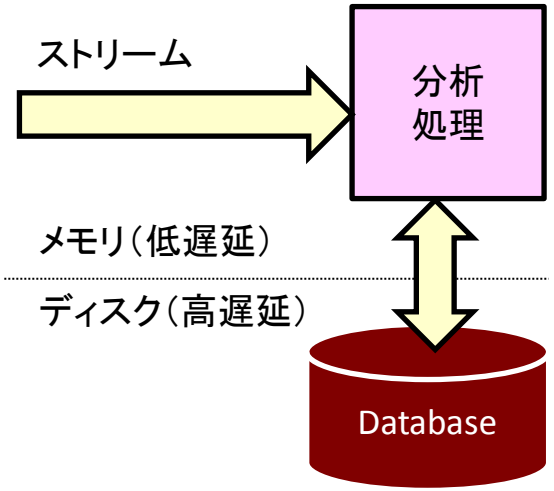
- バッチ処理
  - データが永続的
  - **RDB:** Oracle, Vertica, GreenPlum
  - **NoSQL:** Hadoop, MongoDB

  DBMS: Database Mgmt. System

- リアルタイム処理
  - 問合せが永続的，窓関数
  - **DSMS:** 日立，Sybase, IBM
  - **NoSQL:** STORM, S4

  DSMS: Data Stream Mgmt. System





Part 1: DBMSを用いた浅い分析の報告

データ処理基盤：DBMS

「いないいないばあ」
を含むパケット数
を知りたい.

```
SELECT  COUNT(*)
FROM    packet
WHERE  payload REGEXP /いないいないばあ/
```

20

パケットの
リレーショナルスキーマ

- Destination IP
- Source IP
- Destination Port
- Source Port
- Interface (例：eth0)
- Length (パケット長)
- Version（IPV4 等）
- Payload（パケット内容）

Data

リレーション
packet

Queries are volatile

DBMSの適用

- DBMS
  - PostgreSQL
- スキーマ(右)
- 利用データセット
  - PRACTICE
  - D3M
- 利用情報
  - ヘッダのみ
- データ抽出ツール
  - NEGI

属性名	型	備考
time	text	時間
srcip	text	Source IP
dstip	text	Destination IP
srcport	integer	Source Port
dstport	integer	Destination Port
protocol	integer	プロトコル番号
seqno	bigint	シーケンス番号
ackno	bigint	ACK番号
syn	text	SYNフラグ
ack	text	ACKフラグ
fin	text	FINフラグ
urg	text	URGフラグ
psh	text	PSHフラグ
rst	text	RSTフラグ
szpkt	bigint	IPパケットサイズ
szcontent	bigint	TCPパケットサイズ

ポート番号が同一である通信の検出

対象テーブル　　数え上げ

```
SELECT   srcport, dstport, dstip, COUNT(*)
FROM   practice_dataset_2013_practice_1
WHERE   srcip = '10.220.0.36'
GROUP BY   srcport, dstport, dstip
ORDER BY COUNT(*) DESC LIMIT  5;
```

グルーピング属性

自分のsrcipでフィルタ

PRACTICE 1

srcport	dstport	dstip	count
53555	15510	2.134.49.170	18

Srcportを変えながら同一 dstport へアクセス
（dstipを変える場合もある)

PRACTICE 2

srcport	dstport	dstip	count
68	67	10.220.0.100	324
60432	55003	78.34.188.201	33
59306	55003	78.34.188.201	30
53723	55003	78.34.188.201	30
54299	55003	78.34.188.201	30

PRACTICE 3

srcport	dstport	dstip	count
56138	16471	219.80.142.21	1182
56693	16471	219.80.142.21	1182
52717	16471	136.169.13.7	1148
49313	16471	24.15.221.77	1142
58601	16471	98.145.30.7	1057

PRACTICE 4			
srcport	dstport	dstip	count
61916	16464	189.95.79.14	1149
54793	16464	189.95.79.14	1149
50509	16464	219.55.222.3	1042
63275	16464	219.55.222.3	1042
52628	16464	69.170.93.61	1040

PRACTICE 5			
srcport	dstport	dstip	count
68	67	10.220.0.100	325
58320	4000	89.149.253.239	45
55567	4000	89.149.253.239	36
58579	4000	89.149.253.239	27
60461	4000	89.149.253.239	27

D3M_2012_20120328_malware_20120328_d39569faa627f1aa37ef22ec8			
srcport	dstport	dstip	count
1058	80	217.76.142.78	1401
1060	80	81.31.32.70	1376
1055	80	213.153.32.161	767
1054	53	10.0.0.1	67
1040	8082	186.5.23.154	7

パケットの流量（内から外）

和　　平均　　数え上げ

```
SELECT dstport, SUM(szcontent), AVG(szcontent), COUNT(*)
FROM practice_dataset_2013_practice_1
WHERE srcip = '10.220.0.36'
GROUP BY dstport
ORDER BY SUM DESC LIMIT 5;
```

P1

dstport	sum	avg	count
53	138413	46	3016
25549	104778	177	591
67	102364	314	326
18746	61375	164	373
26614	55162	81	680

P2

dstport	sum	avg	count
55003	114926	9.5	12067
67	102364	314	326
53	15424	45	342
123	496	62	8
80	378	18	21

P3

dstport	sum	avg	count
16471	5789202	8.43	686648
67	29830	314	95
53	4288	45	95
80	336	17	20
123	248	62	4

P4

dstport	sum	avg	count
16464	5319348	14	375582
67	29830	314	95
53	4288	45	95
80	1846	23	80
123	248	62	4

P5

dstport	sum	avg	count
80	222693	37	6202
67	102678	314	327
53	16580	45	362
123	496	62	8
99	0	0	18

D1

dstport	sum	avg	count
8080	50076	46	1089
53	11872	53	224
138	663	221	3

D2

dstport	sum	avg	count
137	10368	64	162
138	663	221	3
53	228	46	5

D3

dstport	sum	avg	count
8080	4917	64	77
8082	4895	64	77
53	4759	71	67
138	1308	218	6
80	731	0.2	3550

多数のアクセスを試みる dstport が存在

参考：TCP handshake検知

```
SELECT COUNT(*)
FROM
   (SELECT *
   FROM $table
   WHERE syn = 'syn' AND protocol = 6) AS syn,
   (SELECT *
   FROM $table
   WHERE ack = 'ack' AND protocol = 6) AS ack
WHERE  syn.srcport = ack.dstport
   AND syn.dstport = ack.srcport
   AND syn.srcip = ack.dstip
   AND syn.dstip = ack.srcip
   AND syn.seqno +1 = ack.ackno;
```

リレーション名	検出数	タプル数	割合(0-1)
D3M_2010_2_pcap_20100308	0	55531	0
D3M_2010_2_pcap_20100309	0	41387	0
D3M_2010_2_pcap_20100311	0	70486	0
D3M_2010_2_pcap_20110208	0	12873	0
D3M_2010_2_pcap_20110214	0	12474	0
D3M_2010_2_pcap_20110216	0	27294	0
practice_dataset_2013_practice_1	3631	54403	0.067
practice_dataset_2013_practice_2	5393	21680	0.249
practice_dataset_2013_practice_3	192706	1160063	0.166
practice_dataset_2013_practice_4	80560	631453	0.128
practice_dataset_2013_practice_5	5999	49137	0.122

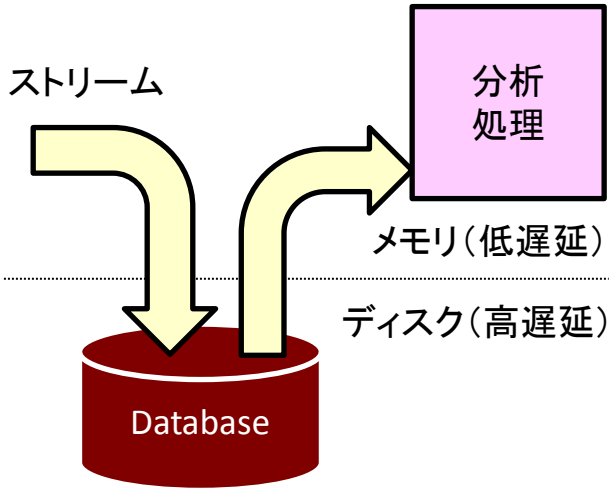
データ処理パラダイム

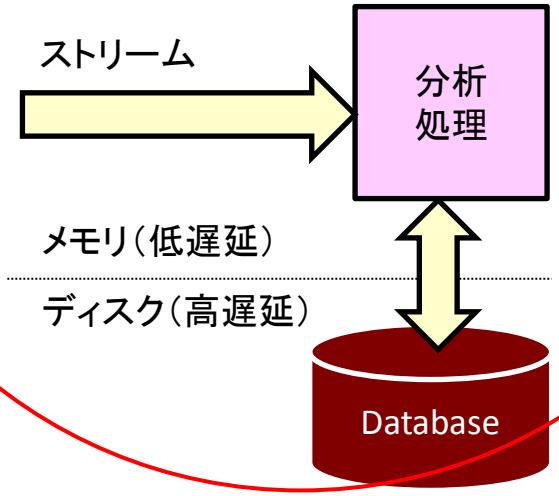
バッチ（DBMS）/ リアルタイム（DSMS）

- バッチ処理
  - データが永続的
  - **RDB:** Oracle, Vertica, GreenPlum
  - **NoSQL:** Hadoop, MongoDB

- リアルタイム処理
  - 問合せが永続的，窓関数
  - **DSMS:** 日立，Sybase, IBM
  - **NoSQL:** STORM, S4

DSMS: Data Stream Processing System





Part 2: 実施予定（3月まで）のDSMS適用の検討

Quick Review: DSMS

「いないいないばあ」
を含むパケット数
を1分ごとに知りたい.

```
SELECT  COUNT(*)
FROM    eth0[TIME 1 MIN, SLIDE 1 MIN]
WHERE  payload REGEXP /いないいないばあ/
```

20

パケットの
リレーショナルスキーマ

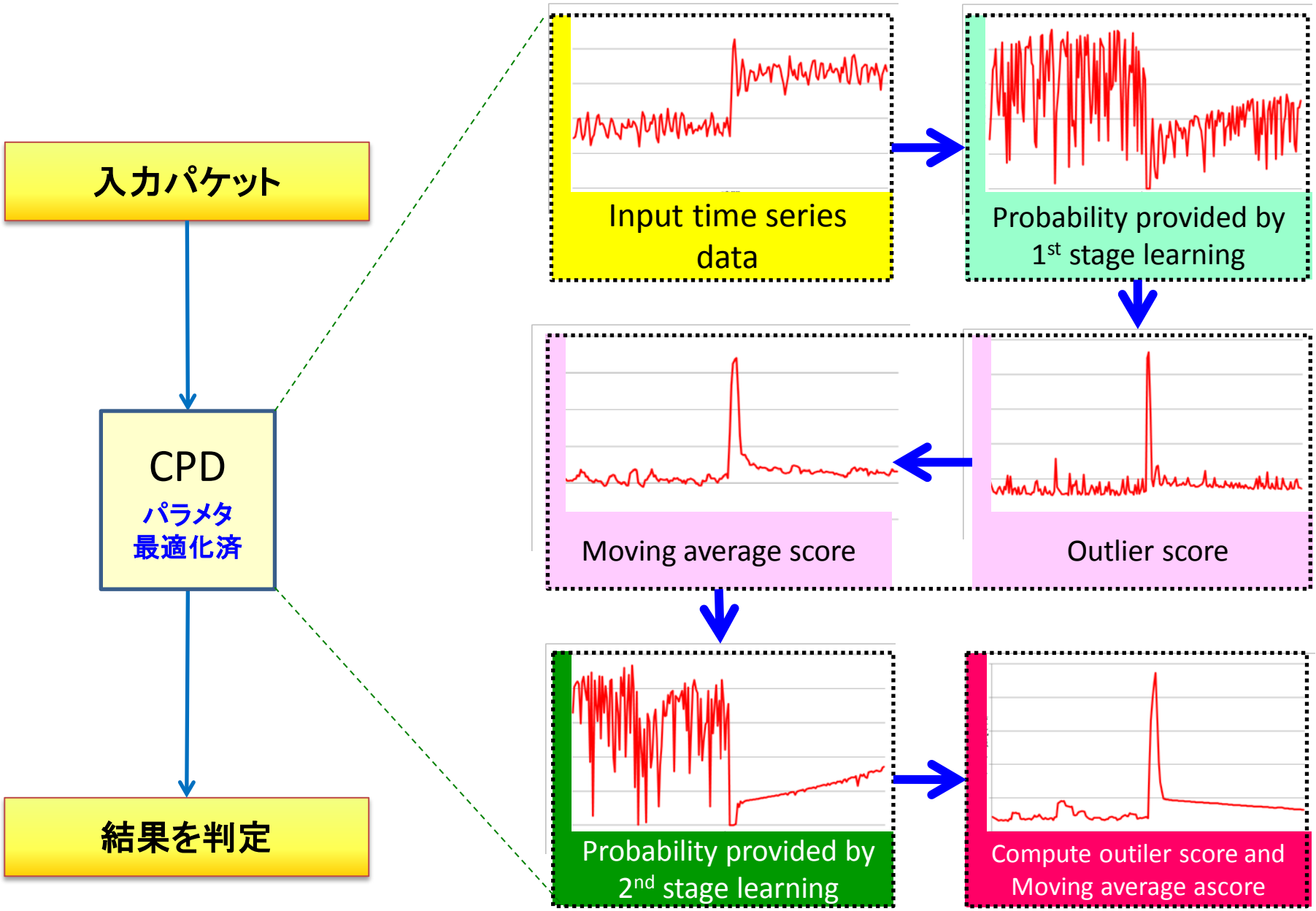
- Destination IP
- Source IP
- Destination Port
- Source Port
- Interface (例：eth0)
- Length (パケット長)
- Version（IPV4 等）
- Payload（パケット内容）

Query

リレーション
eth0

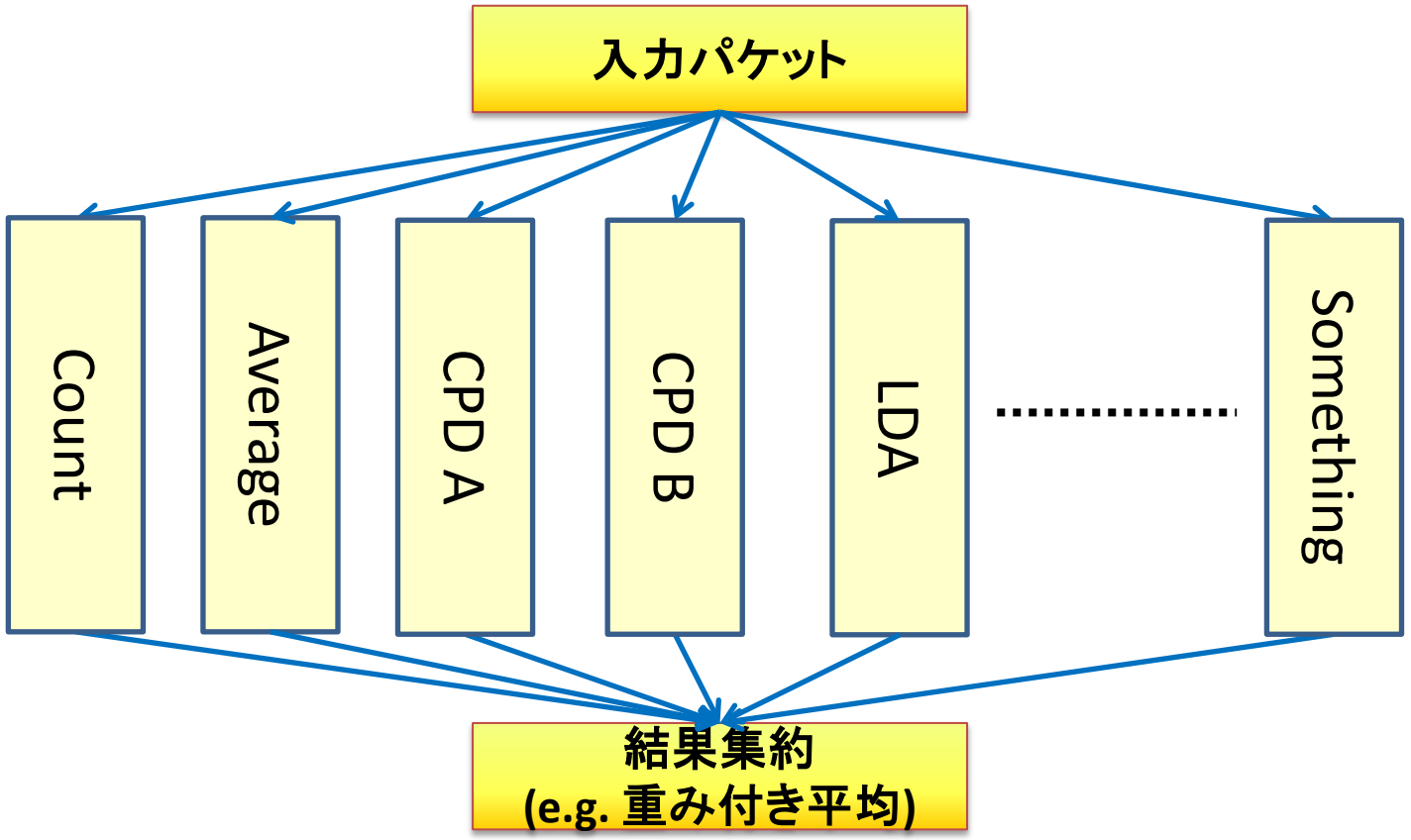
Data are volatile

賢い検知スキーム


入力パケット
CPD
パラメタ
最適化済
結果を判定
Input time series data
Probability provided by 1st stage learning
Moving average score
Outlier score
Probability provided by 2nd stage learning
Compute outiler score and Moving average ascore

力任せなスキーム

多数の手法を同時に適用



管理の必要性-> DSMS: *Falcon* (NEGIを利用)

[NEGI] https://github.com/westlab/negi

再訪：TCP handshake検知

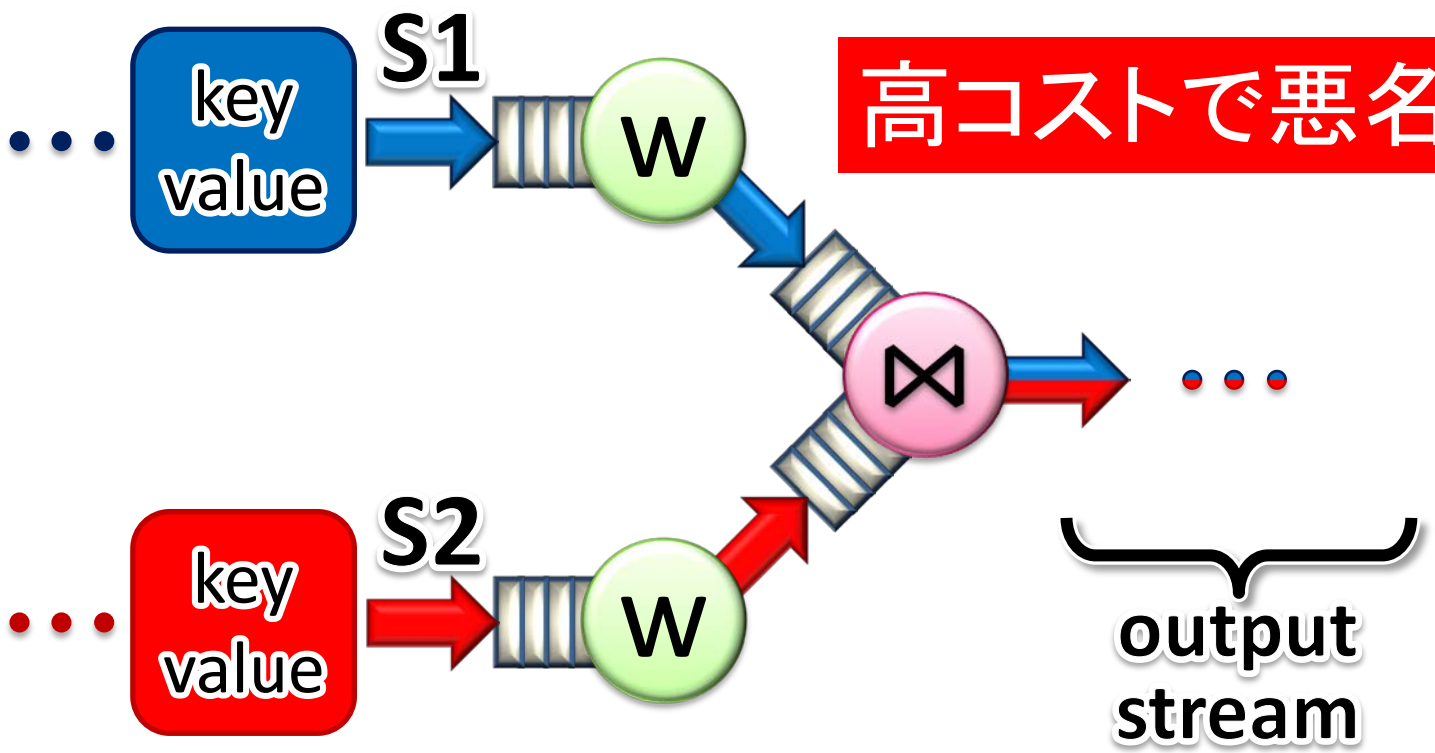
```
SELECT COUNT(*)
FROM
   (SELECT *
   FROM $table
   WHERE syn = 'syn' AND protocol = 6) AS syn,
   (SELECT *
   FROM $table
   WHERE ack = 'ack' AND protocol = 6) AS ack
WHERE  syn.srcport = ack.dstport
   AND syn.dstport = ack.srcport
   AND syn.srcip = ack.dstip
   AND syn.dstip = ack.srcip
   AND syn.seqno +1 = ack.ackno;
```

Join Operation

Join

SELECT	*
FROM	S1 [Rows 100], S2 [Rows 100]
WHERE	S1.key = S2.key

CQL Query



Group-by-aggregate（平均など）: 2億 レコード/秒
@CANDAR'13

Basic: 670万 レコード/秒
Proposal: 1460万 レコード/秒
@SSDBM'13

まとめ

ヘッダ　選択･射影･結合･集約

- 結論：DBMS/DSMSにより，浅い分析は可能
  - Srcportを変えて同一dstport へのアクセス
  - 多数のアクセスを試みる dstport（内→外）

- 課題：深い分析
  - お伺い：どんな手法が必要でしょうか？
  - 準備中の手法
    - Complex Event Processing（A->B)
    - Clustering: Latent Dirichlet Allocation
    - Classification: Support Vector Machne
    - 相関規則発見：Frequent Itemset Mining

謝辞

貴重なデータセットをご用意いただきました
MWS委員会の皆様に心からお礼申し上げます.