

# RFID Data Management

Xiao Pan

## Outline

- [Introduction to RFID technology](#)
- Characteristics of RFID Data
- Research of RFID data management
  - Beginning of the research
  - Fruits of RFID data management
    - Storage and model of RFID
    - Warehousing and Mining Massive RFID Data Sets
    - Data Cleaning
    - Demo
- Conclusion

## What is RFID?

- What is it?
  - RFID(Radio Frequency IDentification) is a technology that allows a **reader** to detect, from a distance, and without line of sight, a unique electronic product code(EPC) that is transmitted by a **tag**.

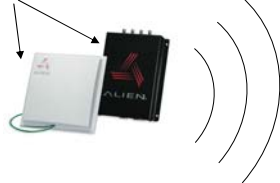
- Tag
  - Attached to items
  - Store item EPC



- Reader
  - Periodical tag scans
  - Records (EPC, time)



Antenna & reader

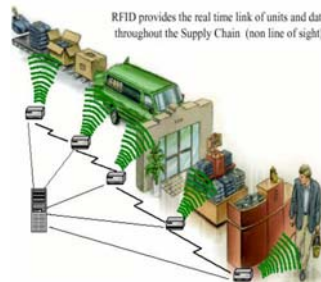


Tags



## Applications

- Supply chain management
  - for example in retail store
- Healthcare
- Airline luggage management
  - (British airways) Implemented to reduce lost/misplaced luggage
- Library
- Something interesting applications
  - CocaCola
  - Fetch money by mobile phone in ATM machine
- ...



# RFID System



## Outline

- Introduction to RFID technology
- **Characteristics of RFID Data**
- Research of RFID data management
  - Beginning of the research
  - Fruits of RFID data management
    - Storage and model of RFID
    - Warehousing and Mining Massive RFID Data Sets
    - Data Cleaning
    - Demo
- Conclusion

# Characteristics of RFID Data

- Large volume
  - A retail with 3000 stores sells 10,000 items a day per store (EPC, location, time)  
Each item 10 traces before leaving store  
How many tuples it will generate each day?  
 $10,000 \times 10 \times 3,000 = 300,000,000$  (without redundancy)
  - Walmart is expected to generate 7 terabytes of RFID data per day

-> model and storage of RFID data
- Inaccurate data
  - Noisy data and duplicate readings

-> Data cleaning of RFID data
- Implicit semantics
  - Observations imply location changes, aggregations, and business processes

-> Query and data mining of RFID data
- Temporal oriented

# Outline

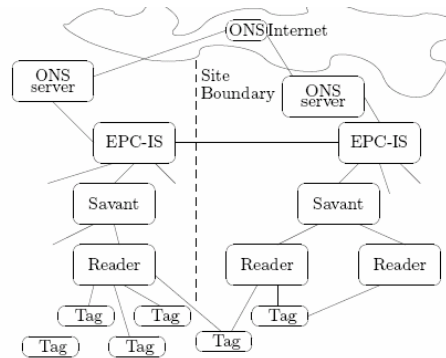
- Introduction to RFID technology
- Characteristics of RFID Data
- Research of RFID data management
  - Beginning of the research
  - Fruits of RFID data management
    - Storage and model of RFID
    - Warehousing and Mining Massive RFID Data Sets
    - Data Cleaning
    - Demo
- Conclusion

# Managing RFID Data\_VLDB2004(invited)

Sudarshan S. Chawathe, Venkat Krishnamurthy etc.

- A layered architecture

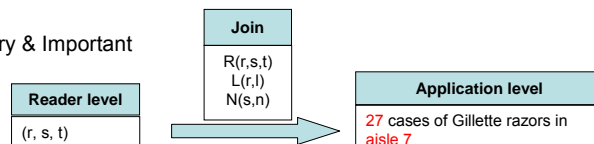
- RFID tags
- Tag readers
- Savant/Middleware
  - Mapping the low-level data stream from readers to a more manageable form that is suitable for application-level interactions
- EPC-IS
  - Most interesting and challenging tasks: Combing business logic with the stream of data emerging from the sensing framework below them
- ONS
  - Essentially a global lookup service



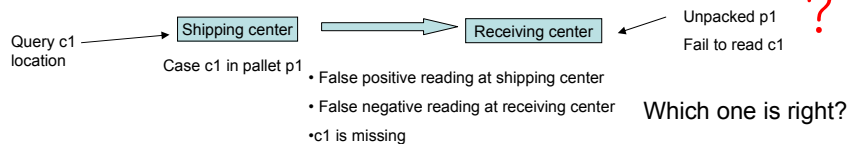
# Managing RFID Data\_VLDB2004(Contd.)

- Inferences

- Necessary & Important



- Challenges: Complex



- RFID data management vs. warehousing

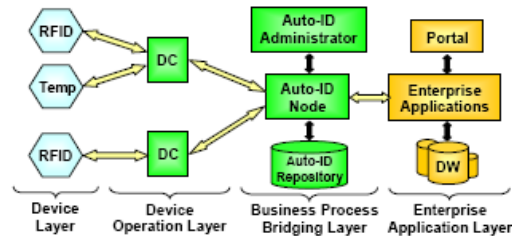
- Analogous task: collecting data, data cleaning etc
- Differences: Currency of data  
 Station-local activities

- Configuration Design

- Determining number, type, and placement of readers, and the manner connected to other sensors
- Design choice affects the amount and nature of data that must be stored at other layers

Integrating Automatic Data Acquisition with Business Processes  
Experiences with SAP's Auto-ID Infrastructure\_VLDB2004(invited)  
Christof Bornhovd, Tao Lin, Stephan Haller, Joachim Schaper

- Auto-ID infrastructure



- Open Issues

- Different Qualities of Service
- Distributed Smart Items Infrastructure
- Seamless Integration of Environmental Sensors
- Privacy

## Outline

- Introduction to RFID technology
- Characteristics of RFID Data
- Research of RFID data management
  - Beginning of the research
  - Fruits of RFID data management
    - [Storage and model of RFID data](#)
    - Warehousing and Mining Massive RFID Data Sets
    - Data Cleaning
    - Demo
- Conclusion

## Supporting RFID-based Item Tracking Applications in Oracle DBMS Using a Bitmap Datatype\_VLDB2005

Ying Hu, Seema Sundara, Timothy Chorma, Jagannathan Srinivasan

- Observation
  - Groups of items in the same proximity - e.g. on a shelf, on a shipment
  - Groups of items with same property - e.g. Same product

- **epc\_bitmap\_segment** Datatype
  - A new type to represent a collection of EPCs with a common prefix

Header 2-bits    EPC\_Manager 21-bits    Object\_Class 17-bits    Serial\_Number 24-bits

0x4AA890001F62C160  
.....  
0x4AA890001FA0B38E

Len	Suff_len	Prefix	Suff_start	Suff_end	bitmap
64	24	0x4AA890001F	0x62C160	0xA0B38E	101001...00010

With EPC Collections

Store_id	Prod_id	Time	Item_collection
s1	p1	t1	epc11, epc12, epc13, ...
s1	p2	t2	epc21, epc22, epc23, ...
...	...	...	...

With epc\_bitmaps

Store_id	Prod_id	Time	Item_bmap
s1	p1	t1	bmap1
s1	p2	t2	bmap2
...	...	...	...

- **Bitmap datatype: multiset of epc\_bitmap\_segment type**  
 CREATE TYPE epc\_bitmap IS  
 TABLE OF epc\_bitmap\_segment;

## Supporting RFID-based Item Tracking Applications in Oracle DBMS Using a Bitmap Datatype (Contd.)

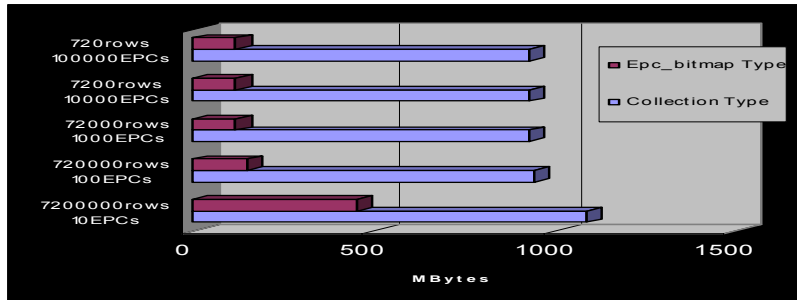
- **epc\_bitmap Operations**
  - Conversion Operations  
epc2Bmap, bmap2Epc, and bmap2Count
  - Pairwise Logical Operations  
bmapAnd, bmapOr, bmapMinus, and bmapXor
  - Maintenance Operations  
bmapInsert and bmapDelete
  - Membership Testing Operation  
bmapExists
  - Comparison Operation  
bmapEqual
- **Use of these operations in SQL**
  - Query: Determine the items added to a shelf between time t1 and t2

Table Shelf\_Inventory

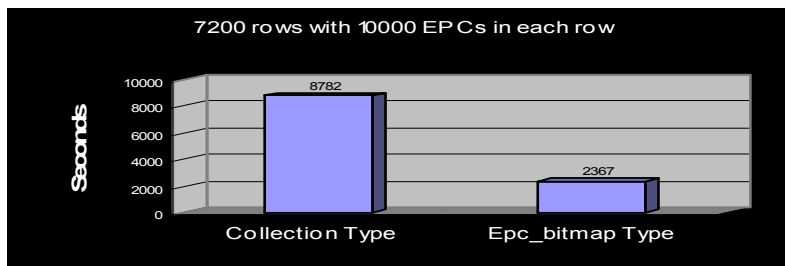
Shelf_id	Time	Item_bmap
sid1	t1	bmp1
sid1	t2	bmp2
...	...	...

```
SELECT bmap2Epc(bmapMinus(s2.item_bmap,s1.item_bmap))
FROM Shelf_Inventory s1, Shelf_Inventory s2
WHERE s1.shelf_id = <sid1> AND
      s1.shelf_id = s2.shelf_id AND
      s1.time=<t1> AND s2.time=<t2>;
```

## Storage Comparison



## Bulk Load Performance



## Temporal Management of RFID Data\_VLDB2005

Fusheng Wang, Peiya Liu

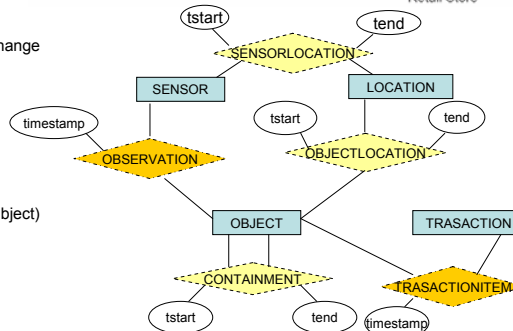
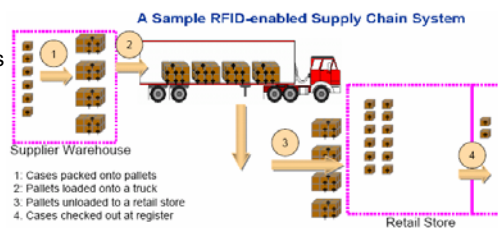
- Dynamic Relationship ER Model (DRER)

- Fundamental Entities in RFID Systems

- Objects
    - Sensors/readers
    - Locations
    - Transactions

- Two dynamic relationships added

- State-based dynamic relationship
      - Object location change
      - Object containment relationship change
      - Reader location change
      - two timestamp attributes : Represents the lifespan of a state
    - Event-based dynamic relationship
      - New event
      - Observations (reader + object)
      - Transacted items (transaction+ object)
      - A timestamp attribute: Represents event occurring time



## Temporal Management of RFID Data\_VLDB2005(Contd.)

- Rules-based RFID Data Transformation



- Rules for location transformation  
`OBSERVATION("r2", e, t) ->`  
`UPDATE:OBJECTLOCATION(e,"L002", t, "UC")`
- Rules for data aggregation  
`seq(s,"r2",Tseq);OBSERVATION("r2", e, t) ->`  
`INSERT:CONTAINMENT(seq(s,"r2",Tseq),e,t,"UC")`
- Rules for data filtering  
`OBSERVATION(Rx, e, Tx), OBSERVATION(Ry, e, Ty),`  
`Rx <> Ry, within(Tx, Ty, T) ->`  
`DROP:OBSERVATION(Rx, e, Tx)`

- Fusheng Wang, Shaorong Liu, Peiya Liu, Bridging Physical and Virtual World: Complex Event Processing for RFID Data Streams, EDBT2006

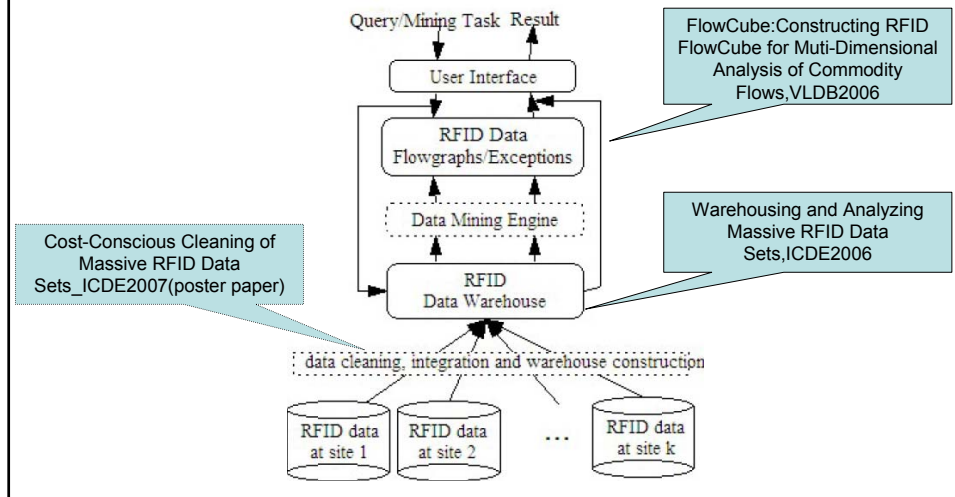
## Outline

- Introduction to RFID technology
- Characteristics of RFID Data
- Research of RFID data management
  - Beginning of the research
  - Fruits of RFID data management
    - Storage and model of RFID data
    - Warehousing and Mining Massive RFID Data Sets
    - Data Cleaning
    - Demo
- Conclusion

## Warehousing and Mining Massive RFID Data Sets\_Keynote for ADMA2006

Jiawei Han

- RFID Warehouse Architecture



## Warehousing and Analyzing Massive RFID Data Sets\_ICDE2006

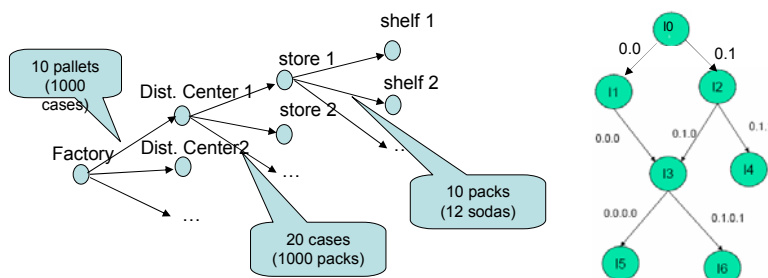
Hector Gonzalez, Jiawei Han, Xiaolei Li, Diego Kabjia

- Why traditional data cube fails?

- View the cleansed RFID data: fact table - (object epc, location, time in, time out : measure).
- measure: count - Number of items that stayed at a given location for a given period.
- Does not consider links within records.
- Example
  - Get the number of items of product  $P$  that traveled from the distribution center  $L$  to stores  $U$ ?
  - We have the count of product  $P$  for each location but we do not know how many of those items went from the first location to the second.
  - Hard to get this information.
  - We need a more powerful model capable of aggregating data while preserving its path-like structure.

## Warehousing and Analyzing Massive RFID Data Sets\_ICDE2006

- Compression Idea: Bulky object movements
  - Objects often **move and stay together** through the supply chain.
  - If 1000 packs of product P stay together at the distribution center : register a single record for all of them.
  - (GID, location, time\_in, time\_out:measures).
  - GID is a generalized identifier that represents the 1000 packs that stayed together at the distribution center



## Warehousing and Analyzing Massive RFID Data Sets\_ICDE2006 (Contd.)

- RFID-Cuboid Architecture
  - **Stay Table:** (GIDs, location, time\_in, time\_out: measures)
    - Records information on items that stay together at a given location
  - **Map Table:** (GID, <GID1,...,GIDn>)
    - Links together stages that belong to the same path. High level GID points to lower level GIDs
    - If saving complete EPC Lists: high costs of IO to retrieve long lists, costly query processing
  - **Information Table:** (EPC list, attribute1,...,attribute n)
    - Records path-independent attributes of the items, e.g., color, manufacturer, price
- Query: Get the number of items of product P that traveled from the distribution center L to stores U?
  - GIDs for L <0.0.0>, <0.1.0>
  - GIDs for U <0.0.0.0>, <0.1.0.1>
  - Prefix pairs: p1: (<0.0.0>, <0.0.0.0>)  
p2: (<0.1.0>, <0.1.0.1>)
  - Retrieve stay records for each pair (including intermediate steps) and compute measure

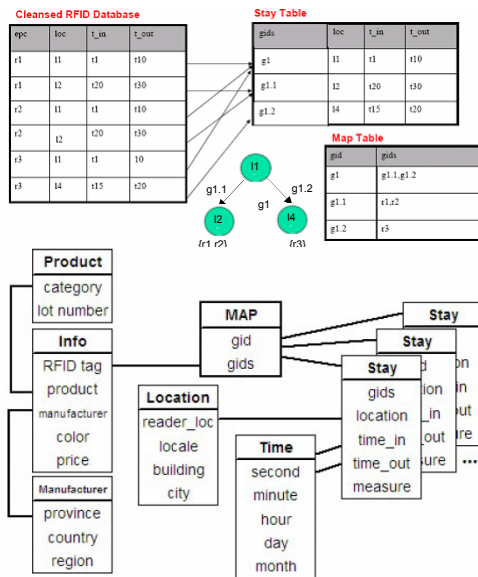


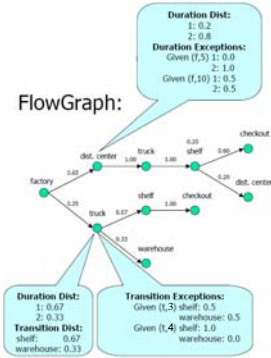
Fig. RFID-Cube Architecture

# FlowCube: Constructing RFID FlowCubes for Multi-Dimensional Analysis of Commodity Flows\_VLDB2006

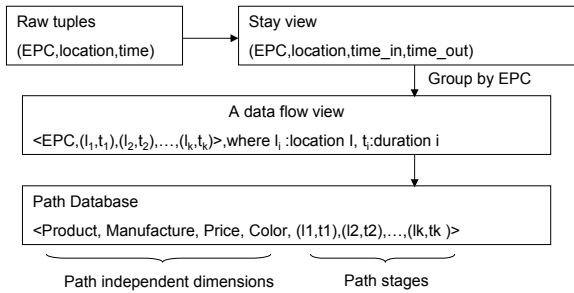
Hector Gonzalez, Jiawei Han, Xiaolei Li

## Data Flow Analysis: FlowGraph

- Tree shaped workflow that summarizes the flow patterns for an item or group of items
  - Nodes: Locations
  - Edges: Transitions
- Each node is annotated with:
  - Distribution of durations at the node
  - Distribution of transition probabilities
  - Exceptions to duration and transition probabilities



## RFID data: A Path Database View



## Path Database:

id	product	brand	path
1	tennis	nike	(f,10)(d,2)(t,1)(s,5)(c,0)
2	tennis	nike	(f,5)(d,2)(t,1)(s,10)(c,0)
3	sandals	nike	(f,10)(d,1)(t,2)(s,5)(c,0)
4	shirt	nike	(f,10)(t,1)(s,5)(c,0)
5	jacket	nike	(f,10)(t,2)(s,5)(c,1)
6	jacket	nike	(f,10)(t,1)(w,5)
7	tennis	adidas	(f,5)(d,2)(t,2)(s,20)
8	tennis	adidas	(f,5)(d,2)(t,3)(s,10)(d,5)

# FlowCube: Constructing RFID FlowCubes for Multi-Dimensional Analysis of Commodity Flows\_VLDB2006(Contd.)

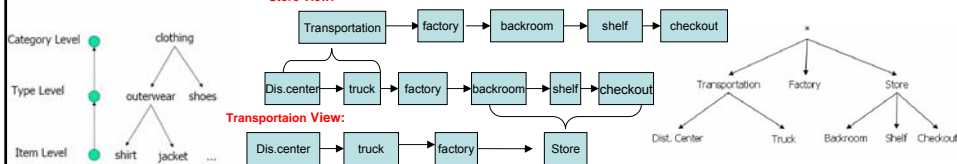
## FlowCube

- Each cuboid has an associated level in the item and path abstraction lattices.
  - Level in the item lattice. (product category, country, price)
  - Level in the path lattice. (<transportation, factory, backroom, shelf, checkout>, hour)

## Path Database:

id	product	brand	path
1	tennis	nike	(f,10)(d,2)(t,1)(s,5)(c,0)
2	tennis	nike	(f,5)(d,2)(t,1)(s,10)(c,0)
3	sandals	nike	(f,10)(d,1)(t,2)(s,5)(c,0)
4	shirt	nike	(f,10)(t,1)(s,5)(c,0)
5	jacket	nike	(f,10)(t,2)(s,5)(c,1)
6	jacket	nike	(f,10)(t,1)(w,5)
7	tennis	adidas	(f,5)(d,2)(t,2)(s,20)
8	tennis	adidas	(f,5)(d,2)(t,3)(s,10)(d,5)

## Product Concept Hierarchy

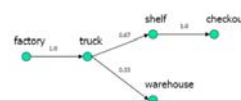


- The measure for each cell in the FlowCube is a FlowGraph computed on the paths aggregated in the cell.

Cuboid for <product type, brand>

cell id	product	brand	path ids
1	shoes	nike	1,2,3
2	shoes	adidas	7,8
3	outerwear	nike	4,5,6

## FlowGraph for cell 3



# Outline

- Introduction to RFID technology
- Characteristics of RFID Data
- Research of RFID data management
  - Beginning of the research
  - Fruits of RFID data management
    - Storage and model of RFID data
    - Warehousing and Mining Massive RFID Data Sets
    - Data Cleaning
    - Demo
- Conclusion

# Issues in Data Cleaning

- False negative reading
  - In this case, RFID tags might not be read by the reader at all while present to a reader
  - Caused by
    - RFID readers capture only 60-70% of all tags that are in the vicinity
    - RF collisions
    - Water or metal shielding
- False positive reading
  - In this case, besides RFID tags to be read, additional unexpected readings are generated
  - Caused by
    - RFID tags outside the normal reading scope of a reader are captured by the reader
    - RFID tags has moved away its vicinity, but reader fails to capture it
    - Unknown reasons from the reader or environment, one of our readers periodically sends wrong IDs
- Duplicate Readings
  - Caused by
    - Tags in the scope of a reader for a long time are read by the reader multiple times
    - Multiple readers are installed to cover larger area or distance, and tags in the overlapped areas read by multiple readers
    - To enhance reading accuracy, multiple tags with same EPCs are attached to the same object, thus generate duplicate readings
- Logical anomalies: tend to be application dependent
  - For example: cycle anomalies
    - (e1, t1, r1, back room)
    - (e1, t1+2, r2, sales floor)
    - (e1, t1+5, r1, back room)
    - (e1, t1+9, r2, sales floor)

## A Pipelined Framework for Online Cleaning of Sensor Data Streams, ICDE 2006(short paper)

Shawn R. Jeffery, Gustavo Alonso, Michael J. Franklin, Wei Hong, Jennifer Widom

- Extensible Sensor stream Processing (ESP):

A declarative query-based framework

- Point

Operates over a single reading in a receptor stream, filtered individual readings(e.g., obvious outliers)

- Smooth

Granularity defined by applications to correct for missed readings temporally (over one input only); uses aggregate function over the input.

- Merge

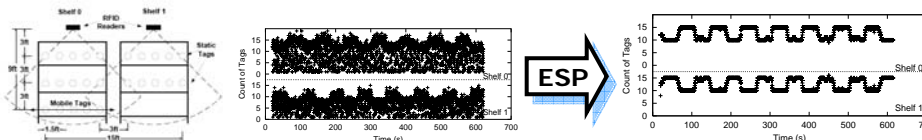
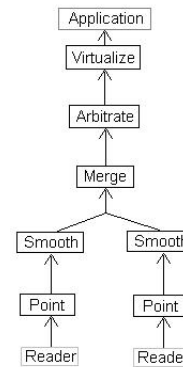
Granularity specified by the application to correct for missed readings spatially

- Arbitrate

Deals with conflicts between different spatial granules

- Virtualize

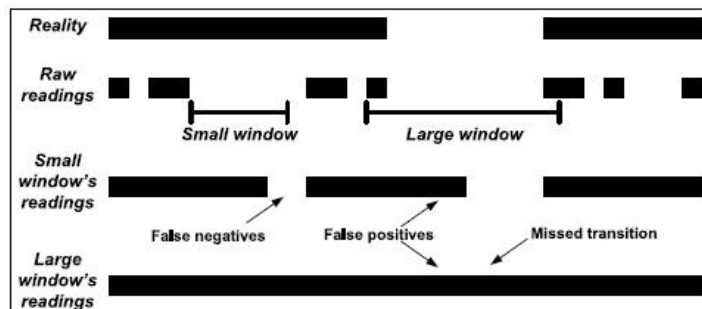
Combine readings from different types of devices and data sources



## Adaptive Cleaning for RFID Data Streams\_VLDB2006

ShawnR. Jeffery, Minos Garofalakis, Michael J.Franklin

- Window Size for RFID Smoothing



- Solution

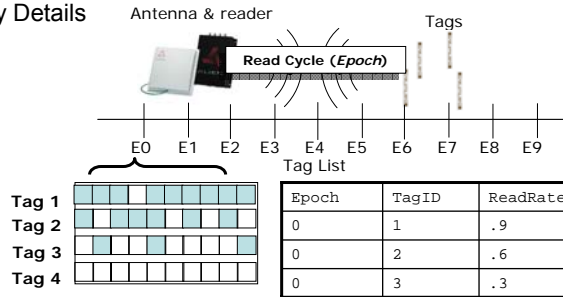
- SMUF(Statistical Smoothing for Unreliable RFID Data)
- Adapt the window size in response to data

## Adaptive Cleaning for RFID Data Streams\_VLDB2006

- Key Insight: A Statistical Sampling Perspective

- RFID data  $\approx$  random sample of present tags,
- Map RFID smoothing to a sampling experiment

- RFID's Gory Details



- RFID Smoothing to Sampling

RFID	Sampling
Read cycle (epoch)	Sample trial
Reading	Single sample
Smoothing window	Repeated trials
Read rate	Probability of inclusion ( $p_i$ )

## Adaptive Cleaning for RFID Data Streams\_VLDB2006

- Per-tag cleaning

- Completeness

$$w_i = \left( \frac{1}{p_i^{avg}} \right) * \ln \left( \frac{1}{\delta} \right)$$

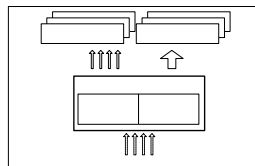
- Transitions

$$\left| |S_{i,j}| - w_i * p_i^{avg} \right| > 2 \sqrt{w_i * p_i^{avg} * (1 - p_i^{avg})}$$

# observed readings # expected readings Is the difference "statistically significant"?

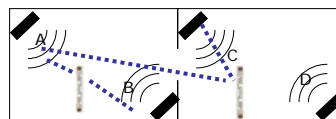
- Mechanisms for:

- Per-tag and multi-tag cleaning



- Not fit for

Two rooms, two readers per room



## Efficiently Filtering RFID Data Streams\_VLDB-CleanDB2006

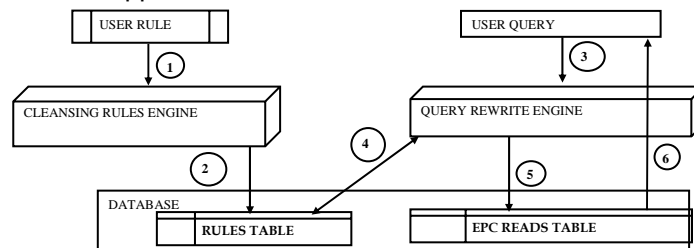
Yijian Bai, Fusheng Wang, Peiya Liu

- Main ideas
  - False positive readings
    - The noise readings are readings with count of distinct tag EPC values below a noise **threshold**
    - Essentially performs the following operations: within any time window with size of **window\_size** ,
      - if the count of the readings with same tag EPC values appears equal to above threshold, then the observed EPC value is not noise and needs to be forwarded for further processing;
      - otherwise the reading is discarded.
  - Duplicate readings
    - If a reading is within **max\_distance** in time from the previous reading with the same key, then this reading is considered a duplicate.
    - Otherwise, it is considered a new reading and is output
- Good points: preserve the original order
- Bad points
  - Not give the cleaning method for false negative reading
  - Don't mention how to confirm the threshold

## A Deferred Cleansing Method for RFID Data Analytics\_VLDB2006

Jun Rao, Sangeeta Doraiswamy, Hetal Thakkar, Latha S.Colby

- Motivation
  - Conventional approach to cleansing is eager
    - Before loading into a warehouse (ETL)
    - Clean once, reuse at query time
    - Typically reducing data size
    - Best strategy if applicable
  - Sometimes eager cleansing is not applicable
    - Don't know how to clean until analyzing the data
    - More than one cleaned version (app-dependant anomalies)
    - Law enforcement (pharmaceutical e-pedigree tracking )
  - Propose deferred cleansing
    - Load everything
    - Clean at query time
    - Has runtime overhead
    - Complementary to eager cleansing
- Overview of This Approach



# Cost-Conscious Cleaning of Massive RFID Data Sets\_ICDE2007(poster paper)

Hector Gonzalez, Jiawei Han, Xuehua Shen

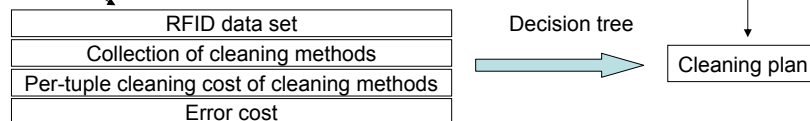
- Motivation

Existing cleaning techniques have focused on the **accurate** methods, but have disregarded the very **high cost** of cleaning in a real application

- Contribution: propose a cleaning framework

Identify the conditions under which a specific **cleaning method** or a **sequence of cleaning methods** should be applied in order to **minimize** the expected **cleaning costs**, including error costs


Input of the framework




## Outline

- Introduction to RFID technology
- Characteristics of RFID Data
- Research of RFID data management
  - Beginning of the research
  - Fruits of RFID data management
    - Storage and model of RFID data
    - Warehousing and Mining Massive RFID Data Sets
    - Data Cleaning
    - Demo
- Conclusion

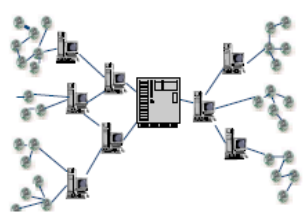
## DEMO(1)



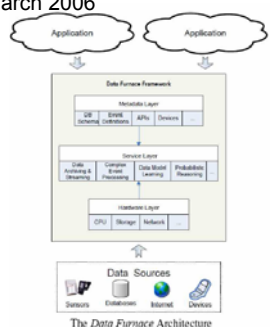
- **HiFi(High Fan-in System)**
  - Database Research Group of UC(University of Californ) Berkeley
  - HiFi is a distributed stream processing system designed to support large-scale sensor-based network
  - Architecture
    - Reader/Sensor
    - Mid-tier devices
    - Host Computer
  - HiFi: A Unified Architecture for High Fan in System, VLDB2004 demo



- **Data Furnace**
  - Intel Research & UC Berkeley
  - Data management for pervasive applications
  - Probabilistic data a first-class citizen
  - Event language-based
  - Probabilistic Data Management for Pervasive Computing: The Data Furnace Project. IEEE Data Engineering Bulletin, Vol. 29, No. 1, March 2006




**Figure 1 - A High Fan-in Architecture**




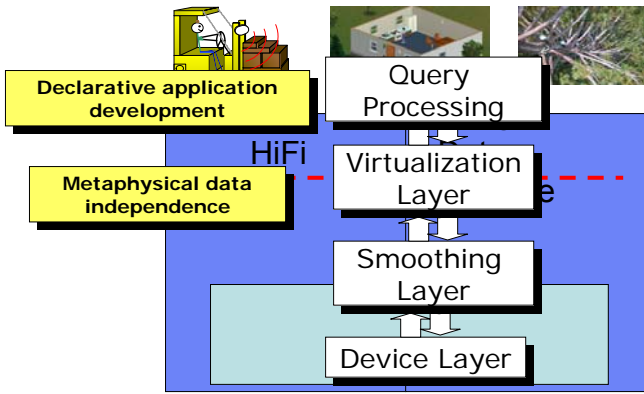
**The Data Furnace Architecture**

## DEMO(1)(Cond.)




$$\begin{matrix} \leftarrow & + & \rightarrow \\ & = & \\ \downarrow & & \downarrow \end{matrix}$$



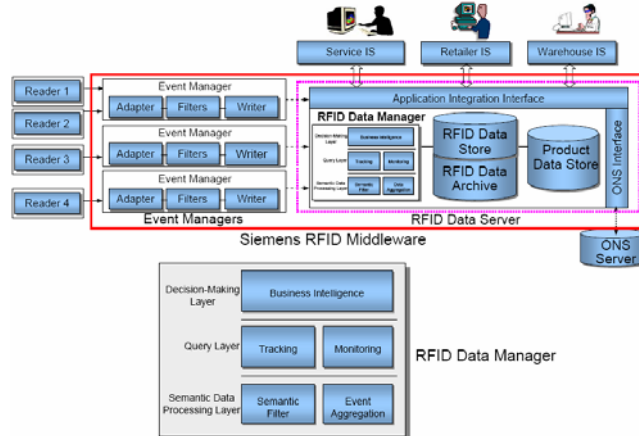


**HiFi**



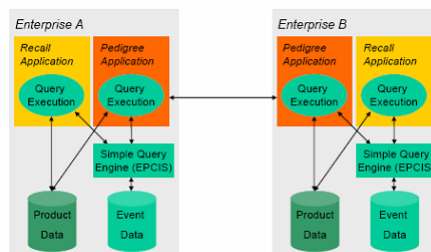
## DEMO(2)

- Simens RFID Middleware
  - Develop and demonstrate at Siemens Corporate Research
  - Applied in health care to increase healthcare safety and workflow efficiency
  - Temporal Management of RFID Data, VLDB2005



## DEMO(3)

- Theseos
  - IBM Almaden Research Center
  - A query engine on top of sovereign, distributed RFID databases, to facilitate traceability query processing
  - Theseos: A Query Engine for Traceability across Sovereign, Distributed RFID Databases, ICDE2007 demo



EPCglobal Approach to Traceability

## Outline

- Introduction to RFID technology
- Characteristics of RFID Data
- Research of RFID data management
  - Beginning of the research
  - Fruits of RFID data management
    - Storage and model of RFID
    - Warehousing and Mining Massive RFID Data Sets
    - Data Cleaning
    - Demo
- Conclusion

## Conclusion

- What is RFID?
- Some fruits of research on RFID data management
  - Storage and model of RFID data
  - Warehousing and Mining Massive RFID Data Sets
  - Data cleaning of RFID data
  - Existing demo
    - HiFi and Data Furnace
    - Simens RFID Middleware
    - Theseos
- What can we do??