RFID Analytics

Cleaning, Warehousing, and Mining RFID Data

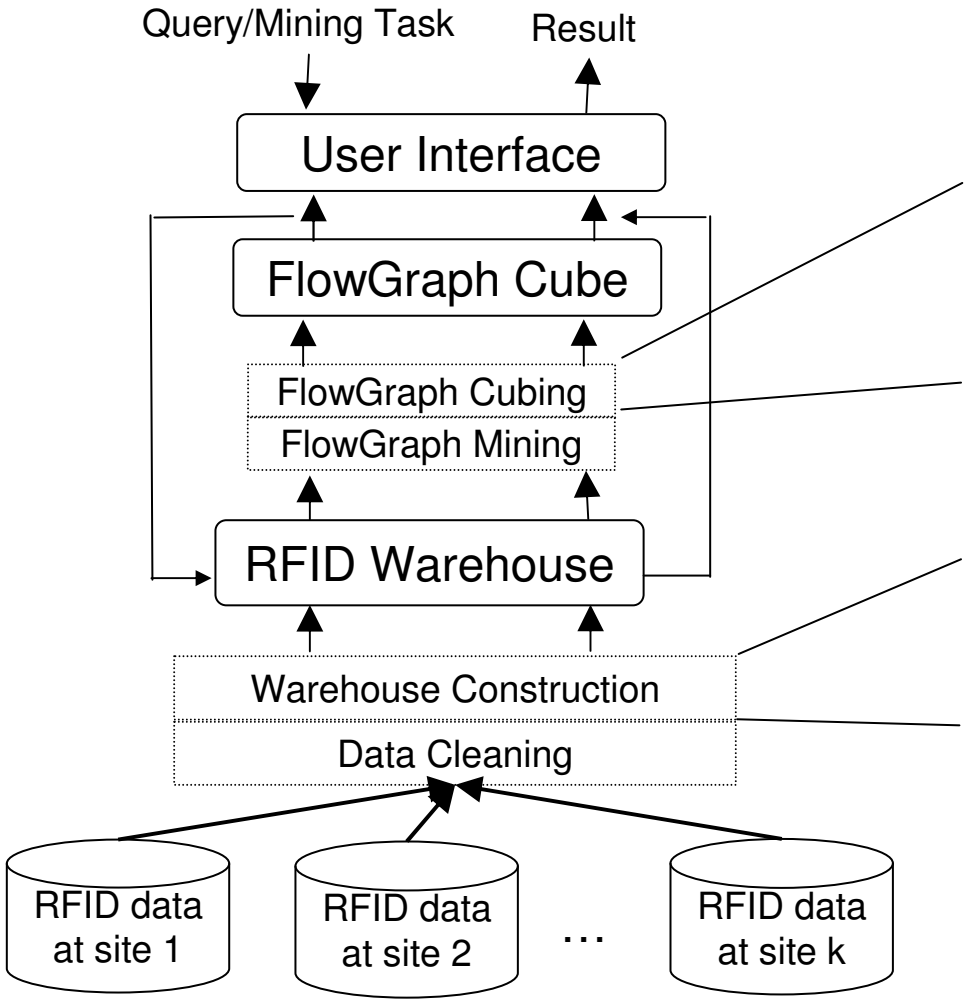
by

Hector Gonzalez

Advisor
Jiawei Han

University of Illinois at Urbana Champaign

Overview



FlowCube: Constructing RFID FlowCubes for Multi-Dimensional Analysis of Commodity Flows [Gonzalez, Han, Li. VLDB'06]

Mining Compressed Commodity Workflows From Massive RFID Data Sets [Gonzalez, Han, Li. CIKM'06]

Warehousing and Analyzing Massive RFID Data Sets [Gonzalez, Han, Li. ICDE'06]

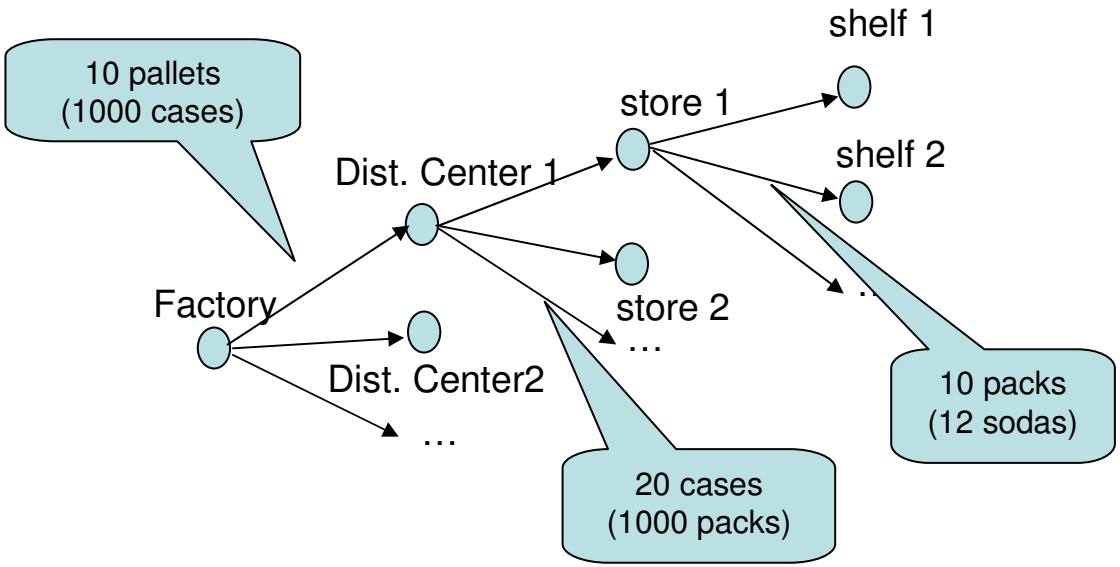
Cost-Conscious Cleaning of Massive RFID Data Sets [Gonzalez, Han, Shen, Klabjan. ICDE'07]

Warehousing

Why Warehousing?

- Huge data sets, terabytes generated each day
- We need OLAP to make sense of the data
- Traditional data cubes don't work. A data cube only provides aggregates for a given combination of dimension values. We need aggregates at the path level.
  - What is the average temperature of milk when moving from Farm A to Warehouse W and finally to Store B?

Compression Ideas

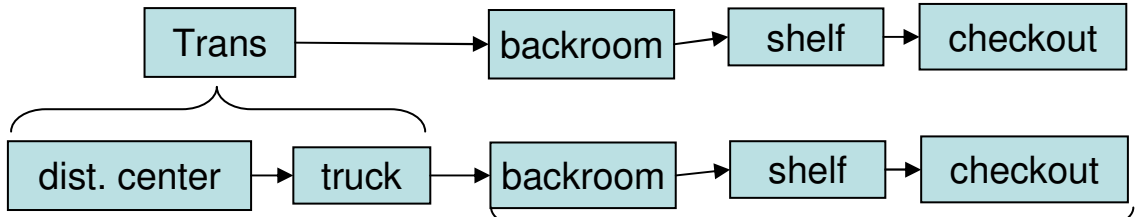


Bulky Movements

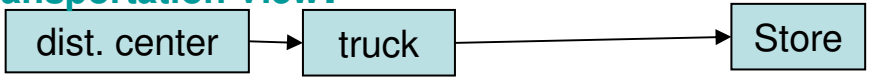
- Products move in large groups through supply chain
- Compress EPC lists into a group of generalized identifiers (GIDs)

Path Generalization

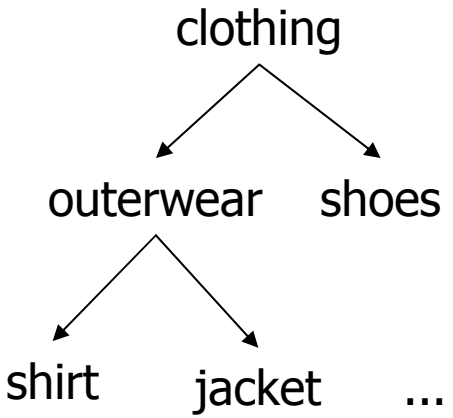
Store View:



Transportation View:

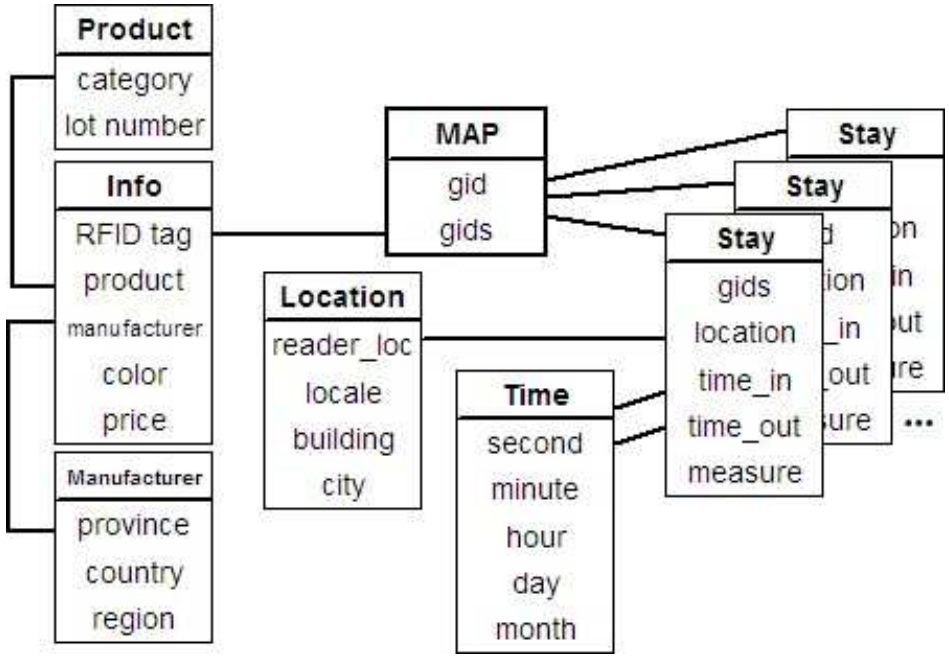


Item Generalization

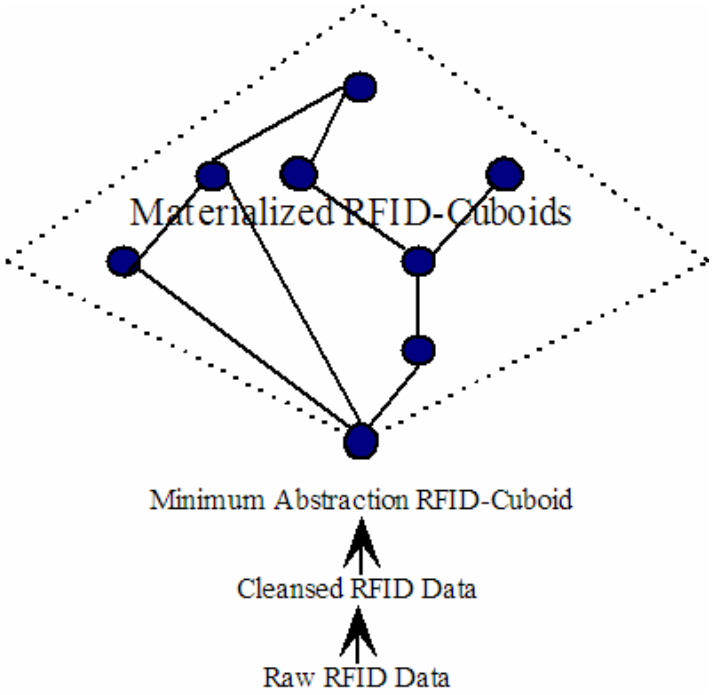


RFID Cube

RFID Cuboid



Materialized Cuboids



Query Processing

- Support for OLAP: roll-up, drill-down, slice, and dice
- Path query: New to RFID-Warehouses, **about the structure of paths**
  - What products that go through quality control have shorter paths?
  - What locations are common to the paths of a set of defective auto-parts?
  - Identify containers at a port that have deviated from their historic paths

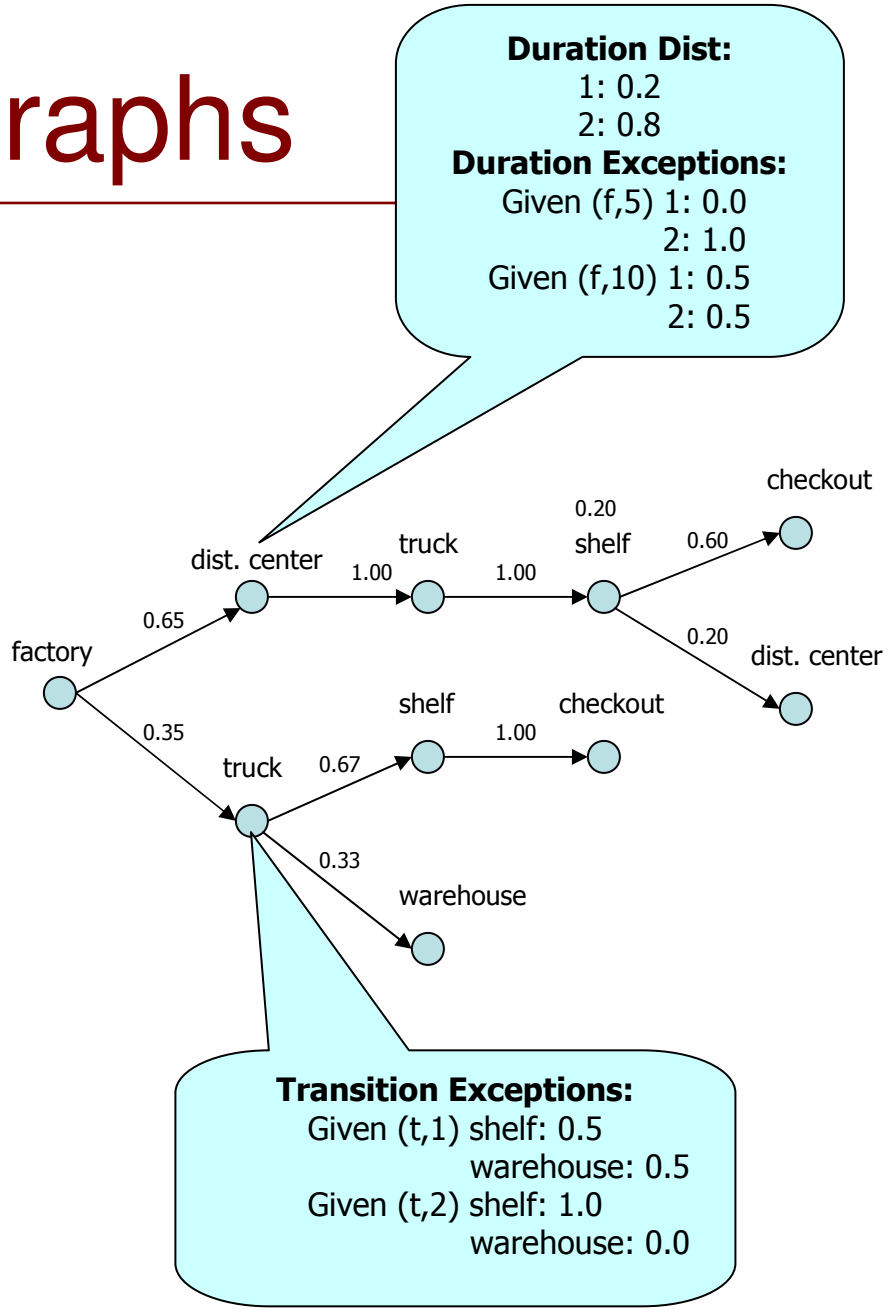
FlowGraph Mining

Why FlowGraphs?

- Compact summary of popular paths traversed by items
- Highlights important deviations from popular paths
  - Laptops stay at the shelf for 3 weeks on average, unless they stayed for more than 1 month in the warehouse, in which case stay increases to 2 months.
  - Milk transitions from the refrigerator to the checkout counter with probability 90%, unless it stayed at non-refrigerated locations in which case it will transition to dumpster with probability 80%.

FlowGraphs

- Tree shaped workflow
- Captures main trends and significant deviations
- Answers to
  - Is there a correlation between time at quality control and returns?
  - What do paths of items discarded in Boston have in common?
  - List containers arriving at NY that present unusual paths.



FlowGraph Cubing

- Data Cube where each cell is a FlowGraph.
- Multi-dimensional summaries of flow patterns at different abstraction levels bring insight into flow trends, exceptions, evolution, and can be used to optimize business processes.
- The FlowCube goes beyond the traditional Data Cube with scalar aggregates, and adds a path view of the data.

Cleaning

Why Cleaning?

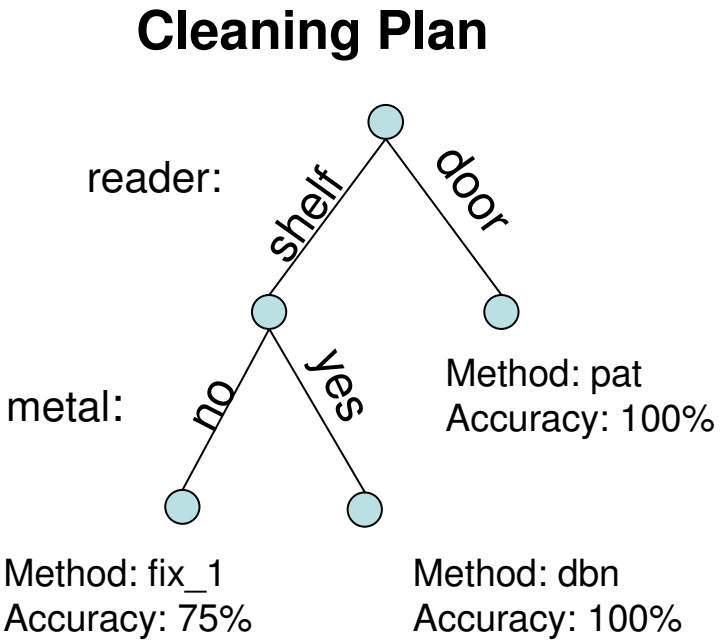
- RFID data in inherently dirty (50% + loss rates)
  - Large number of false negatives
    - RF Interference caused by water, metal, or RF signals cause dropped readings
    - Moving tags, or tags positioned far away from reader are usually missed.
  - False positives
    - Readers detect items that should not really be read, e.g. Item goes near door reader and it is detected but only items that go through door should be detected.

Why cost-conscious cleaning?

- The volume of RFID data is enormous, we need to track thousands of items per second. Cleaning speed is essential.
- How to clean such enormous volume of information in real time? Medical testing analogy:
  - Diagnose a disease (item location) using the least expensive medical test (cleaning method) that is effective, escalate to more expensive tests only when absolutely necessary

Cleaning Plan

- Define a cost model:
  - Assign costs to cleaning methods (training cost, execution cost, maintenance cost, etc).
  - Define error costs.
- Using training data determine the efficacy of training methods under different contexts.
- Optimization Problem: Construct a cleaning plan (methods to apply for different circumstances) that minimizes the total expected cleaning costs.



Thank You