Announcements

• HW3 due on Nov 30 (Fri), 5 pm
• Final report due on Dec 12, Wed
• Class presentation next lecture (Thurs, Nov 29)
  — In the order of your group no.
  — 4 min presentation
  — See details from the announcements in the last lecture
• Please fill out the course evaluations!
  — We need to hear from each of you
  — We need your feedback/suggestions to keep improving this class

Reading Material

• [RG]
  — Chapter 25
• Gray-Chaudhuri-Reynolds-Verdoliva-Fall-Pirahesh, ICDE 1996 “Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals”
• Harinarayan-Rajaraman-Ullman, SIGMOD 1996 “Implementing data cubes efficiently”

Acknowledgement:
• The following slides have been created adapting the instructor material of the [RG] book provided by the authors Dr. Ramakrishnan and Dr. Gehrke.
• Some slides have been prepared by Prof. Shivnath Babu

Data Warehousing

• Growing industry: $8 billion way back in 1998
• Data warehouse vendor like Teradata
  • big “Petabyte scale” customers
  • Apple, Walmart (2008-2.5PB), eBay (2013-primary DW 9.2 PB, other big data 40PB, single table with 1 trillion rows), Verizon, AT&T, Bank of America
  • supports data into and out of Hadoop
• Lots of buzzwords, hype
  — slice & dice, rollup, MOLAP, pivot, ...

Introduction

• Organizations analyze current and historical data
  — to identify useful patterns
  — to support business strategies
• Emphasis is on complex, interactive, exploratory analysis of very large datasets
• Created by integrating data from across all parts of an enterprise
• Data is fairly static
• Relevant once again for the recent “Big Data analysis”
  — to figure out what we can reuse, what we cannot

### OLTP vs. Data Warehousing/OLAP

<table>
<thead>
<tr>
<th>OLTP</th>
<th>Data Warehousing/OLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mostly updates</td>
<td>Mostly reads</td>
</tr>
<tr>
<td>Applications: Order entry, sales update, banking transactions</td>
<td>Applications: Decision support in industry/organization</td>
</tr>
<tr>
<td>Detailed, up-to-date data</td>
<td>Summarized, historical data (from multiple operational db, grows over time)</td>
</tr>
<tr>
<td>Structured, repetitive, short tasks</td>
<td>Query intensive, ad hoc, complex queries</td>
</tr>
<tr>
<td>Each transaction reads/updates only a few tuples (tens of)</td>
<td>Each query can accesses many records, and perform many joins, scans, aggregates</td>
</tr>
<tr>
<td>MB-GB data</td>
<td>GB-TB data</td>
</tr>
<tr>
<td>Typically clerical users</td>
<td>Decision makers, analysts as users</td>
</tr>
</tbody>
</table>

#### Important:
- Consistency, recoverability, Maximizing tr. throughput
- Important: Query throughput, Response times

---

### Three Complementary Trends

- **Data Warehousing (DW):**
  - Consolidate data from many sources in one large repository
  - Loading, periodic synchronization of replicas
  - Semantic integration

- **OLAP:**
  - Complex SQL queries and views.
  - Queries based on spreadsheet-style operations and “multidimensional” view of data.
  - Interactive and “online” queries.

- **Data Mining:**
  - Exploratory search for interesting trends and anomalies

---

### ROLAP and MOLAP

- **Relational OLAP (ROLAP)**
  - On top of standard relational DBMS
  - Data is stored in relational DBMS
  - Supports extensions to SQL to access multi-dimensional data

- **Multidimensional OLAP (MOLAP)**
  - Directly stores multidimensional data in special data structures (e.g. arrays)

---

### ROLAP: Star Schema

- To reflect multi-dimensional views of data
- Single fact table
- Single table for every dimension
- Each tuple in the fact table consists of
  - pointers (foreign key) to each of the dimensions (multi-dimensional coordinates)
  - numeric value for those coordinates
- Each dimension table contains attributes of that dimension
- No support for attribute hierarchies

---

### Dimension Hierarchies

- For each dimension, the set of values can be organized in a hierarchy:

<table>
<thead>
<tr>
<th>PRODUCT</th>
<th>TIME</th>
<th>LOCATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>category</td>
<td>year</td>
<td>year</td>
</tr>
<tr>
<td>name</td>
<td>quarter</td>
<td>quarter</td>
</tr>
<tr>
<td></td>
<td>country</td>
<td>country</td>
</tr>
<tr>
<td>week</td>
<td>month</td>
<td>month</td>
</tr>
<tr>
<td></td>
<td>state</td>
<td>state</td>
</tr>
<tr>
<td>date</td>
<td>city</td>
<td>city</td>
</tr>
</tbody>
</table>

---

### ROLAP: Snowflake Schema

- Refines star-schema
- Dimensional hierarchy is explicitly represented
- (+) Dimension tables easier to maintain
  - suppose the “category description” is being changed
- (-) Need additional joins
- Fact Constellations
  - Multiple fact tables share some dimensional tables
  - e.g. Projected and Actual Expenses may share many dimensions

---

Duke CS, Fall 2018
CompSci 516: Database Systems
OLAP and Data Cube

Motivation: OLAP Queries
- Data analysts are interested in exploring trends and anomalies
  - Possibly by visualization (Excel) - 2D or 3D plots
  - "Dimensionality Reduction" by summarizing data and computing aggregates
- Find total unit sales for each
  1. Model
  2. Model, broken into years
  3. Year, broken into colors
  4. Year
  5. Model, broken into color, ...

Sales (Model, Year, Color, Units)

Roll-Ups
- Analysis reports start at a coarse level, go to finer levels
- Order of attribute matters
- Not relational data (empty cells no keys)

<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Color</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chevy</td>
<td>1994</td>
<td>Black</td>
<td>50</td>
</tr>
<tr>
<td>Chevy</td>
<td>1994</td>
<td>White</td>
<td>40</td>
</tr>
<tr>
<td>Chevy</td>
<td>1995</td>
<td>Black</td>
<td>115</td>
</tr>
<tr>
<td>Chevy</td>
<td>1995</td>
<td>White</td>
<td>85</td>
</tr>
</tbody>
</table>

• Roll-ups are asymmetric

Data Cube: Intuition

More complex to do these with GROUP-BY

Data Cube
- Computes the aggregate on all possible combinations of group by columns.
  - If there are N attributes, there are 2^N-1 super-aggregates.
  - If the cardinality of the N attributes are C_1,..., C_N then there are a total of (C_1+1)...(C_N+1) values in the cube.
- ROLL-UP is similar but just looks at N aggregates
Data Cube

- SQL Server

```
SELECT Model, Year, Color, sum(units)
FROM Sales
GROUP BY Model, Year, Color
WITH CUBE
```

Data Cube Syntax

Basic Ideas

- Need to compute all group-by-s:
  - ABCD, ABC, ABD, BCD, AB, AC, AD, BC, BD, CD, A, B, C, D

- Compute GROUP-BYs from previously computed GROUP-BYs
  - e.g. first ABCD
  - then ABC or ACD
  - then AB or AC ...

- Which order ABCD is sorted, matters for subsequent computations
  - if (ABCD) is the sorted order, ABC is cheap, ACD or BCD is expensive

Implementing Data Cube

Notations

- ABCD
  - group-by on attributes A, B, C, D
  - no guarantee on the order of tuples

- (ABCD)
  - sorted according to A > B > C > D

- ABCD and (ABCD) and (BCDA)
  - all contain the same results
  - but in different sorted order

Optimization 1: Smallest Parent

- Compute GROUP-BY from the smallest (size) previously computed GROUP-BY as a parent
  - AB can be computed from ABC, ABD, or BCD
  - ABC or ABD better than ABCD
  - Even ABC or ABD may have different sizes, try to choose the smaller parent

Ack: from slides by Laurel Orr and Jeremy Hyrkas, UW
Optimization 2: Cache Results

• Cache result of one GROUP-BY in memory to reduce disk I/O
  – Compute AB from ABC while ABC is still in memory

Optimization 3: Amortize Disk Scans

• Amortize disk reads for multiple GROUP-BYS
  – Suppose the result for ABCD is stored on disk
  – Compute all of ABC, ABD, ACD, BCD simultaneously in one scan of ABCD

Optimization 4: Shared partition

• for hash-based algorithms
  – Uses hash tables to compute smaller GROUP-BYS
  – If the hash tables for AB and AC fit in memory, compute both in one scan of ABC
  – Otherwise partition on A, and compute HTs of AB and AC in different partitions
• “pipe-hash” algorithm not covered in class

Optimization 5: Shared-sort

• From one sorted order of (ABCD)
  – Compute (ABC)
  – Compute (AB)
  – Compute (A)
  – Compute “all”

PipeSort: Shared-sort optimization

• BUT, may have a conflict with “smallest-parent” optimization
  – (ABD) \( \Rightarrow \) (AB) could be a better choice
  – Figure out the best parent choice by running a weighted-matching algorithm layer by layer (details not covered in class)
Data Mining

Reading Material

Optional Reading:
1. [RG]: Chapter 26

- 23,863 citations on Google Scholar in November 2018
  - 23,038 in November 2017
  - 20,610 in November 2016
  - 19,496 in April 2016

One of the most cited papers in CS!

Acknowledgement:
The following slides have been prepared adapting the slides provided by the authors of [RG] and using several presentations of this paper available on the internet (esp. by Ofer Pasternak and Brian Chase)

Duke CS, Fall 2018
CompSci 516: Database Systems

Four Main Steps in KD and DM (KDD)

- **Data Selection**
  - Identify target subset of data and attributes of interest

- **Data Cleaning**
  - Remove noise and outliers, unify units, create new fields, use denormalization if needed

- **Data Mining**
  - Extract interesting patterns

- **Evaluation**
  - Present the patterns to the end users in a suitable form, e.g. through visualization

Duke CS, Fall 2018
CompSci 516: Database Systems

Several DM/KD (Research) Problems

- **Discovery of causal rules**
- **Learning of logical definitions**
- **Fitting of functions to data**
- **Clustering**
- **Classification**
- **Inferring functional dependencies from data**
- **Finding “usefulness” or “interestingness” of a rule**

  - See the citations in the Agarwal-Srikant paper
  - Some discussed in [RG] Chapter 27

Duke CS, Fall 2018
CompSci 516: Database Systems

Mining Association Rules

- **Retailers collect and store massive amounts of sales data**
  - Transaction date and list of items

- **Association rules:**
  - E.g. 98% customers who purchase “tires” and “auto accessories” also get “automotive services” done
  - Customers who buy mustard and ketchup also buy burgers
  - Goal: find these rules from just transactional data (transaction id + list of items)

Duke CS, Fall 2018
CompSci 516: Database Systems

Applications

- **Can be used for**
  - Marketing program and strategies
  - Cross-marketing (mass e-mail, webpages)
  - Catalog design
  - Add-on sales
  - Store layout
  - Customer segmentation
Notations

- Items $I = \{i_1, i_2, \ldots, i_m\}$
- $D$: a set of transactions
- Each transaction $T \subseteq I$
  - has an identifier $TID$
- Association Rule
  - $X \rightarrow Y$ (not Functional Dependency!)
  - $X \cap Y = \emptyset$

Confidence and Support

- Association rule $X \rightarrow Y$
- Confidence $c = |\text{Tr. with } X \text{ and } Y| / |\text{Tr. with } X|$
  - $c\%$ of transactions in $D$ that contain $X$ also contain $Y$
- Support $s = |\text{Tr. with } X \text{ and } Y| / \text{all Tr.|}$
  - $s\%$ of transactions in $D$ contain $X$ and $Y$.

Support Example

<table>
<thead>
<tr>
<th>TID</th>
<th>Cereal</th>
<th>Beer</th>
<th>Bread</th>
<th>Bananas</th>
<th>Milk</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Support(Cereal)
  - $4/8 = 5$
- Support(Cereal $\rightarrow$ Milk)
  - $3/8 = 37.5$

Confidence Example

<table>
<thead>
<tr>
<th>TID</th>
<th>Cereal</th>
<th>Beer</th>
<th>Bread</th>
<th>Bananas</th>
<th>Milk</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Confidence(Cereal $\rightarrow$ Milk)
  - $3/4 = 75$
- Confidence(Bananas $\rightarrow$ Bread)
  - $1/3 = 33.3\%$

$X \rightarrow Y$ is not a Functional Dependency

For functional dependencies

- F.D. = two tuples with the same value of of $X$ must have the same value of $Y$
  - $X \rightarrow Y \Rightarrow XZ \rightarrow Y$ (concatenation)
  - $X \rightarrow Y \rightarrow Z \Rightarrow X \rightarrow Z$ (transitivity)

For association rules

- $X \rightarrow A$ does not mean $XY \rightarrow A$
  - May not have the minimum support
  - Assume one transaction $(A)$
- $X \rightarrow A$ and $A \rightarrow Z$ do not mean $X \rightarrow Z$
  - May not have the minimum confidence
  - Assume two transactions $(A)(A)$

Problem Definition

- Input
  - a set of transactions $D$
    - Can be in any form – a file, relational table, etc.
  - min support (minsup)
    - min confidence (minconf)
- Goal: generate all association rules that have
  - support $>=$ minsup and
  - confidence $>=$ minconf
Decomposition into two subproblems

- 1. Apriori
  - for finding "large" itemsets with support >= minsup
  - all other itemsets are "small"
- 2. Then use another algorithm to find rules $X \rightarrow Y$ such that
  - Both items $X \cup Y$ and $X$ are large
  - $X \rightarrow Y$ has confidence >= minconf
- Paper focuses on subproblem 1
  - if support is low, confidence may not say much
  - subproblem 2 in full version of the paper

Basic Ideas - 1

- Q. Which itemset can possibly have larger support: ABCD or AB
  - i.e. when one is a subset of the other?
- Ans: AB
  - any subset of a large itemset must be large
  - So if AB is small, no need to investigate ABC, ABCD etc.

Basic Ideas - 2

- Start with individual (singleton) items (A, B) ...
- In subsequent passes, extend the "large itemsets" of the previous pass as "used"
- Generate new potentially large itemsets (candidate itemsets)
- Then count their actual support from the data
- At the end of the pass, determine which of the candidate itemsets are actually large
  - becomes seed for the next pass
- Continue until no new large itemsets are found

Notations

- Assume the database is of the form <TID, I1, I2, ...> where items are stored in lexicographic order
- TID = identifier of the transaction
- Also works when the database is "normalized": each database record is <TID, item> pair

Algorithm Apriori

- $L_k = \{\text{large} \ (k-\text{items})\}$
- For ($k = 2$; $L_{k-1} \neq \emptyset$) do begin
  - $C_k = \text{aprior-gen}(L_{k-1}, \text{Min})$
  - for all transactions $t$ do begin
    - $C_k \text{subsets} \subset C_k$
  - count candidates $c \in C_k$ do $c\text{-count} += \text{mincount}$
  - $L_k = \{c \in C_k | \text{count} \geq \text{mincount}\}$
  - end
  - end
  - $\text{Answer} = \bigcup_k L_k$

Apriori-Gen

- Takes as argument $L_k$, (the set of all large $k$-itemsets)
- Returns a superset of the set of all large $k$-itemsets by augmenting $L_k$

  - Join step
    - $L_k$ with $L_{k-1}$
    - $p$ and $q$ are two large ($k$-itemsets)
    - Add $p$ and $q$ into $C_k$
    - where $P : \cup \{\text{item} \text{ }_{\text{of }} L_k, p, q\}$
    - Join by adding the last item of $q$ to $p$

  - Prune step
    - for all subsets $c \subset C_k$ do
      - for all ($k$-itemsets) $s$ of $c$ do
        - if $(s \subseteq L_k)$ then
          - delete $c$ from $C_k$
    - Check all the subsets, remove all candidate with some "small" subset
Apriori-Gen Example - 1

Step 1: Join \( k = 4 \)

Assume numbers 1-5 correspond to individual items

\[
\begin{align*}
L_3 & \quad C_4 \\
\{1,2,3\} & \quad \{1,2,3,4\} \\
\{1,2,4\} & \quad \{1,2,4\} \\
\{1,3,4\} & \quad \{1,3,4\} \\
\{1,3,5\} & \quad \{1,3,5\} \\
\{2,3,4\} & \quad \{2,3,4\}
\end{align*}
\]

Apriori-Gen Example - 2

Step 1: Join \( k = 4 \)

Assume numbers 1-5 correspond to individual items

\[
\begin{align*}
L_3 & \quad C_4 \\
\{1,2,3\} & \quad \{1,2,3,4\} \\
\{1,2,4\} & \quad \{1,2,3,4\} \\
\{1,3,4\} & \quad \{1,3,4,5\} \\
\{1,3,5\} & \quad \{1,3,5\} \\
\{2,3,4\} & \quad \{2,3,4\}
\end{align*}
\]

Apriori-Gen Example - 3

Step 2: Prune \( k = 4 \)

- Remove itemsets that can’t have the required support because there is a subset in it which doesn’t have the level of support i.e. not in the previous pass \( k-1 \)

\[
\begin{align*}
L_3 & \quad C_4 \\
\{1,2,3\} & \quad \{1,2,3,4\} \\
\{1,2,4\} & \quad \{1,2,3,4\} \\
\{1,3,4\} & \quad \{1,3,4,5\} \\
\{1,3,5\} & \quad \{1,3,5\} \\
\{2,3,4\} & \quad \{2,3,4\}
\end{align*}
\]

Correctness of Apriori

- Insert into \( C_k \)
- Select \( p, q, r, \ldots, q' \)
- From \( L_{k-1} \)
- Where \( p \cup \{q, r, \ldots, q'\} \) has required support

Check yourself

- Show that \( C_k \supset L_k \)
- Any subset of large itemset must also be large
- For each \( p \) in \( L_k \), it has a subset \( q \) in \( L_{k-1} \)
- We are extending those subsets \( q \) in \( L_{k-1} \) with another subset \( q' \) of \( p \), which must also be large
- Equivalent to extending \( L_{k-1} \) with all items and removing those whose \( (k-1) \) subsets are not in \( L_{k-1} \)
- Prune is not deleting anything from \( L_k \)

Take-Aways

- DBMS Basics
- DBMS Internals
- Overview of Research Areas
- Hands-on Experience in DB systems

Conclusions

(of 516, Fall 2017)
DB Systems

- Traditional DBMS
  - PostGres, SQL
- Large-scale Data Processing Systems
  - Spark/Scala, AWS
- New DBMS/NOSQL
  - MongoDB

- In addition
  - XML, JSON, JDBC, Python/Java

DB Basics

- SQL
- RA/logical Plans
- RC
- Datalog
  - Why we needed each of these languages
- Normal Forms

DB Internals and Algorithms

- Storage
- Indexing
- Operator Algorithms
  - External Sort
  - Join Algorithms
- Cost-based Query Optimization
- Transactions
  - Concurrency Control
  - Recovery

Large-scale Processing and New Approaches

- Parallel DBMS
- Distributed DBMS
- Map Reduce
- NOSQL

Other Topics

- Data Warehouse/OLAP/Data Cube
- Association Rule Mining

- Hope some of you will further explore
  Database Systems/Data Management/Data Analysis/Big Data!