Advanced Topics in Modern Databases, Big Data Processing

Sunnie Chung
Data Warehousing and OLAP

- Introduction
  - Decision Support Technology
  - On Line Analytical Processing
- Star Schema
- Relational Aggregation Operators
  - Data Cube
  - Roll Up, Drill Down
- Papers to Cover
  - An Overview of Data Warehousing and OLAP Technology by Surajit Chaudhuri (Microsoft) and Umeshwar Dayal (HP Labs), *in the proceedings of IEEE 1995*
  - Data Cube: A Relational Aggregation Operator Generalizing Group By, Cross Tab, and SubTotals by Jin Gray (Microsoft), et al, *in the proceedings of IEEE 1996*
Data Mining

- Introduction
  - Overview of Data Mining
  - Data Cleaning Process
- Main Concepts
  - Association Rules
  - Support and Confidence
- APRIORI Algorithm
- Optimization Techniques of Apriori Algorithm
- Example
  - Pandora
- Book to Cover
  - J. Han, M. Kamber (2001), *Data Mining*, Morgan Kaufmann Publishers, San Francisco, CA
Semistructured Data vs Structured Data

- **Structured Data**
  - Relational Database, Data Warehouse
  - SQL

- **Semistructured Data**
  - XML, HTML
  - XQuery, XPath

- **Unstructured Data**
  - Text, Web Data (Mixed of Text, Image, Audio, Video)

- **Problems and Solutions of Processing Semistructured/Unstructured Data**

- **Introduction of Mark Up Languages:**
  - XML, HTML

- **Papers To Cover (to come)**
Advanced XML

- Introduction of XML
- XML Schema, Semantics, Protocol
- XQuery, XPath
- XML Query Processing
- Papers To Cover
  - XML 1.0 (DOM and SAX2 APIs)
  - XQuery 1.0 and XPath 2.0 Semantics
  - XSLT
Information Retrieval: How Google Search Engine Works

- Web data Processing
- Building Relevancy Metric for Search Engine
- Example:
  - Google Search Engine
- Papers to Cover (to come)
Map Reduce and Apache Hadoop

- Introduction of Google Map Reduce
- Introduction of Apache Hadoop
- HDFS Architecture
- Parallel programming and the MapReduce programming model
- Algorithms of Map Reduce/Apache Hadoop
- Papers to Cover
  - MapReduce: Simplified Data Processing on Large Clusters by Jeffrey Dean (Google) and Sanjay Ghemawat (Google) *in the proceedings of OSDI 2004*
    - [http://www.cs.vu.nl/~ralf/MapReduce/paper.pdf](http://www.cs.vu.nl/~ralf/MapReduce/paper.pdf)
  - Open Source MapReduce
  - Apache Hadoop *in White Papers by Apache, Yahoo*
Data Warehouse on Parallel Processing

- Architecture of Parallel Processing
- Flow of Data Processing in Data Warehouse on Parallel Processing Architecture
  - Retrieval
  - Aggregation
- Example:
  - Teradata Parallel Architecture
  - Oracle Parallel Architecture
Business Intelligence

- What is Business Intelligence??
- Data Mining on Web Transaction Data of Business
- Upcoming Big Trend for Every Business for Current and Next Generation
Big Data: Web Data Processing

- Problems of Big Data Processing
- Solutions
- Processing Web Transaction Data (Unstructured Data) on Data Warehouse (Structured Data Processing Server)
Real World Examples

- Face Book
- Papers to Cover
  - Data Warehousing and Analytics Infrastructure at Facebook by Ashish Thusoo (Facebook), *et al*, *in the proceedings of Sigmod 2010*
- Ebay
  - Data Warehousing and Analytics at Ebay
Data Warehouse that Supports Unstructured Web Data

- Distributed data warehouse that supports unstructured content and more aspects of extreme data
- ETL: External Transfer Language
- SQL-MapReduce
  - SQL language extension using ETL and Stored Procedure that transfers Unstructured Data (Web Data) to Structured Data (Relation: Table)
- Processing the Converted Relational Database on 100% Parallel Shared Nothing Distributed System
Cloud Computing

• Introduction of Cloud Computing
  Public Cloud
  Private Cloud
  Hybrid Cloud

• Architecture
  100% Shared Nothing Parallel System

• Example
  • Amazon Cloud