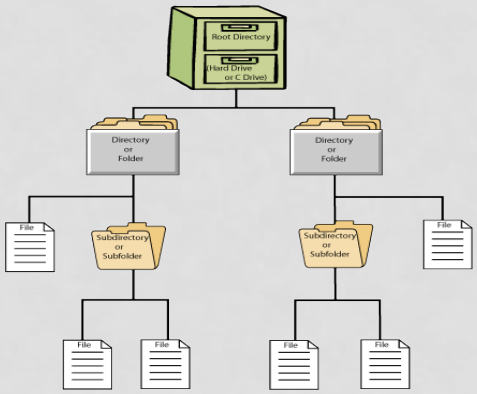
CS561- ADVANCED TOPICS IN DATABASE SYSTEMS

CS561-SPRING 2012
WPI, MOHAMED ELTABAKH

INTRODUCTION & LOGISTICS

HISTORY OF DBMS

- Database systems have evolved since 70s to replace the file system w.r.t storing and querying the data



File system

DBMS

WHY DBMS ???

Storing and querying the data in file system has many disadvantages

- **Data redundancy and inconsistency**
  - Multiple file formats, duplication of information in different files
  - Multiple records formats within the same file
  - No order enforced between fields
- **Difficulty in accessing data**
  - Need to write a new program to carry out each new task
  - No indexes, always scan the entire file
- **Integrity problems**
  - Modify one file (or field in a file), and not changing the dependent fields or files
  - Integrity constraints (e.g., account balance > 0) become "buried" in program code rather than being stated explicitly

WHY DBMS (CONT'D) ???

- **Concurrent access by multiple users**
  - Many users need to access/update the data at the same time (***concurrent access***)
  - Uncontrolled concurrent access can lead to inconsistencies
    - Example: Two people are updating the same bank account at the same time
- **Security problems**
  - Hard to provide user access to some, but not all, data
- **Recovery from crashes**
  - While updating the data the system crashes
- **Maintenance problems**
  - Hard to search for or update a field
  - Hard to add new fields

DBMS PROVIDES SOLUTIONS

- **Modeling of applications semantics and constraints**
- **Data consistency even with multiple users**
- **Efficient access to the data**
- **Data integrity embedded in the DBMS**
- **Recovery from crashes, security**

TRADITIONAL APPLICATIONS OF DBMS

- Transactional data, banking systems, retail stores, airline reservations, restaurant systems, etc…

- **Characteristics of these applications**
  - Simple and well-structured data
  - No complex relationships or operations
  - Simple data types
  - Querying and reporting is not very complex

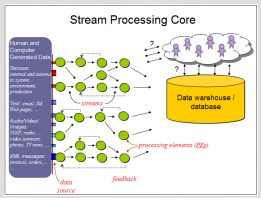
Given these ingredients ➔ Relational Database Systems (RDBMS) is a perfect system

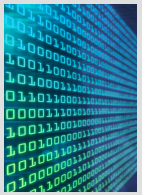
EMERGING APPLICATIONS !!!

- DBMSs are the natural home of the data
  - Because of all DBMSs desired properties
- But, applications are getting **more complex**
  - The assumed characteristics of simplicity no longer hold
- Database management systems have to change and expand to cope with the new requirements and challenges
- Tons of research on advanced topics in DBMSs in many directions
  - New data models and data formats
  - New features and access methods
  - New optimizations and query processing
  - …

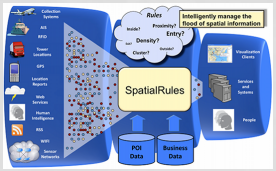
EXAMPLES OF EMERGING APPLICATIONS

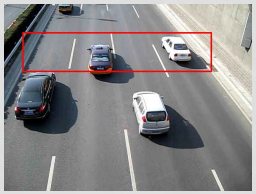
- **Data Stream Management Systems**
  - Data are continuously arriving (no persistency)
  - One-pass main memory processing
  - Load balancing and load shedding





- **Moving objects and spatio-temporal applications**
  - Continuous streams of moving objects
  - Data, by definition, has two key dimensions (space & time)
  - Special query types, e.g., range queries, KNN queries

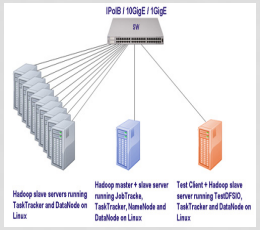


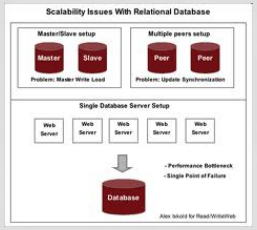


EXAMPLES OF EMERGING APPLICATIONS

- **Scientific Data Management**
  - E.g., in biology, chemistry, physics, atmospheric science, etc.
  - Complex data types, e.g., arrays, images, sequences, structures
  - Metadata, annotations and comments about the data
  - Complex processing and workflows
  - Provenance and lineage information

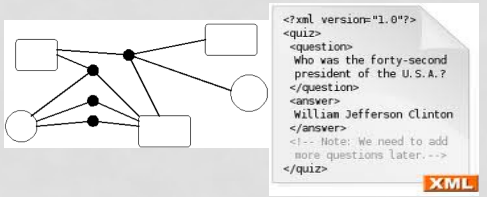
- **Large-Scale Data Analytics and Distributed Processing**
  - Massive scale data processing (terabytes and petabytes)
  - Highly distributed and parallel processing
  - New infrastructure and computing paradigms
  - Distributed DBMSs and Hadoop/MapReduce framework



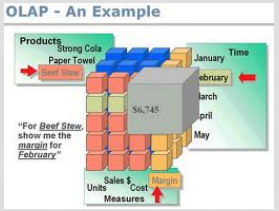


EXAMPLES OF EMERGING APPLICATIONS

- **Data Models for Complex Structures**
  - Object-oriented data model (OODBMS)
  - Object-relational data model (ORDBMS)
  - Semi-structured data model (XML)



- **Data Integration and Data Mining/OLAP**
  - Integrating data from various sources
  - Entity resolution, schema mapping, etc.
  - Discovering hidden knowledge (without the users knowing what they want)



The list goes on and on….

COURSE PLAN AND ROADMAP

- Touch various advanced topics in database systems
- **Lectures will have two flavors**
  - **Typical presentations** (given by the instructor) covering book chapters
  - **Research-oriented presentations** (given by students) covering research papers

COURSE PLAN AND ROADMAP (WHAT YOU EXPECT TO LEARN)

- **Typical presentations will cover (By instructor)**
  - Object-oriented and object-relational data models
  - Semi-structured (XML) data model
  - Distributed and parallel database
  - Active Databases and authorizations
  - Information Integration and OLAP
  - Hadoop and scientific data management

50% of lectures

- **Research-oriented presentations (By students)**
  - **Flexibility based on your interest**
  - **Suggested areas are:**
    - Scientific data management
    - Hadoop/MapReduce Infrastructure
    - Keyword search in database systems
    - Cloud computing
    - Data integration

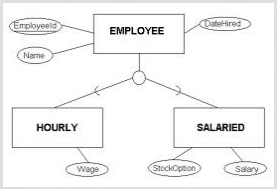
50% of lectures

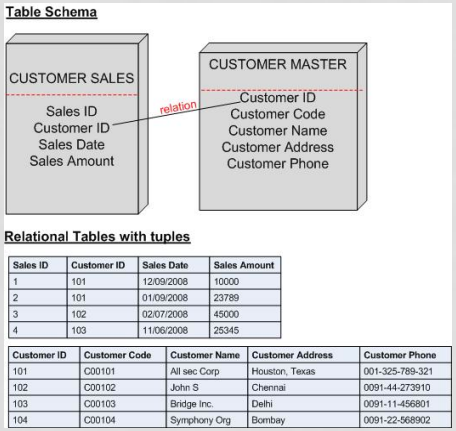
BRIEF OVERVIEW ON COURSE'S TOPICS

(Typical Presentations)

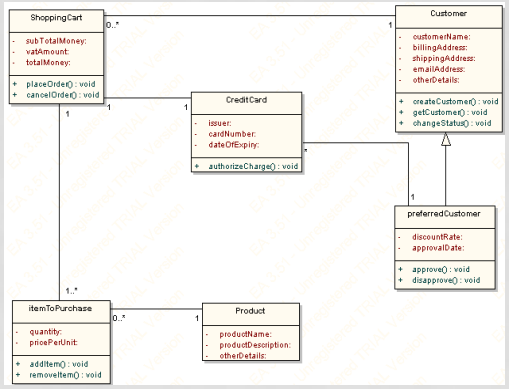
1- OBJECT-ORIENTED & OBJECT-RELATIONAL MODEL

- Relations are the key concept, everything else is around relations
- Primitive data types, e.g., strings, integer, date, etc.
- Great normalization, query optimization, and theory
- **Application are getting more complex**
  - CAD: Computer Aided Design, CAM: Computer aided manufacture
  - Multimedia, document management, telecommunication
- **What is missing in relational model ??**
  - Handling of complex objects and complex relationships
  - Handling of complex data types
  - Code is not coupled with data
  - No inherence, encapsulation, etc.





Relational model



Object-Oriented model

1- OBJECT-ORIENTED & OBJECT-RELATIONAL MODEL

- **Object-Oriented Database (OODBMS)**
  - Depends purely on concepts from OO programming, e.g., C++ or Java
  - Define classes, objects, inheritance, etc.
  - Tries to take some concepts from the relational model, e.g., SELECT statement
  - New languages ODL (object definition language) & OQL (object query language)

```
1)  class Movie {
2)     attribute string title;
3)     attribute integer year;
4)     attribute integer length;
5)     attribute enum Film {color,blackAndWhite} filmType;
    };
```

```
SELECT m.year
FROM Movies m
WHERE m.title = "Gone With the Wind"
```

ODL & OQL

- **Object-Relational Database (ORDBMS)**
  - Still the fundamental concept is 'Relation'
  - Extend the relational model with concepts from OO programming, e.g., complex types, inherence, encapsulation, etc.
  - Extended SQL called SQL3 (or SQL-99)

```
1)  CREATE TYPE MovieType AS (
2)       title    CHAR(30),
3)       year     INTEGER,
4)       inColor BOOLEAN
    );
```

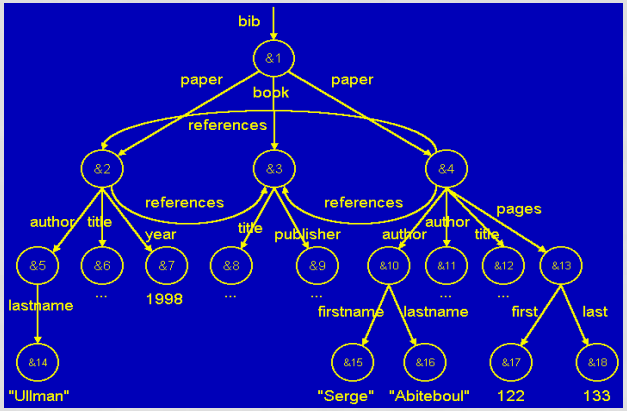
```
5)  CREATE TABLE Movie OF MovieType (
6)       REF IS movieID SYSTEM GENERATED,
7)       PRIMARY KEY (title, year)
    );
```

SLQ-99

	No Query	Query
Complex Data	OODBMS	ORDBMS
Simple Data	File System	RDBMS

2-SEMISTRUCTURED (XML) DATA MODEL

- **Key motivation is the flexibility**
  - Schema is not fixed or not known in advance
  - New attributes or optional attributes
  - Different cardinality for different objects
- **Other models have schema, but semi-structured model is *schemaless***
  - Data is self-describing through the ***tagging*** system
- **XML has two modes**
  - Well-formed XML ---No Schema at all
  - Valid XML --- governed by DTD (Document Type Definition)
    - More flexible than relational or OO models
    - Allows validation and more optimizations and pre-processing



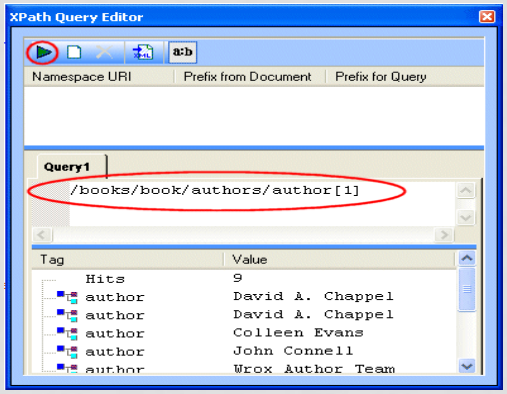
Semi-structured model (Tree—without relationships, Graph—with relationships)

```
<Books>
    <Book ISBN="0553212419">
        <title>Sherlock Holmes: Complete Novels...
        <author>Sir Arthur Conan Doyle</author>
    </Book>
    <Book ISBN="0743273567">
        <title>The Great Gatsby</title>
        <author>F. Scott Fitzgerald</author>
    </Book>
    <Book ISBN="0684826976">
        <title>Undaunted Courage</title>
        <author>Stephen E. Ambrose</author>
    </Book>
    <Book ISBN="0743203178">
        <title>Nothing Like It In the World</title>
        <author>Stephen E. Ambrose</author>
    </Book>
</Books>
```

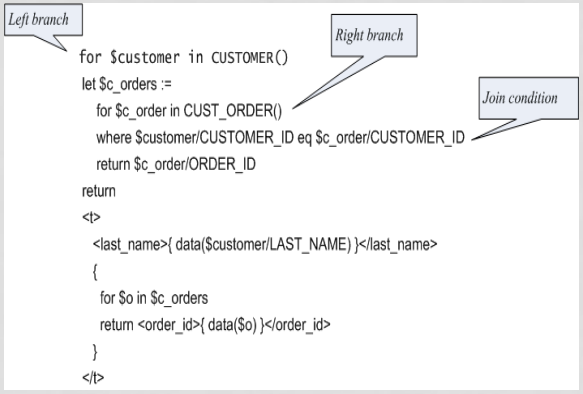
XML document

2-SEMISTRUCTURED (XML) DATA MODEL

- **Programming and Query Languages**
  - **XPath:** Path expressions to navigate in a graph of semi-structured data
  - **XQuery:** extension to XPath by adopting features from SQL
  - **XSLT:** document transformation to produce another XML document or HMTL document



XPath example

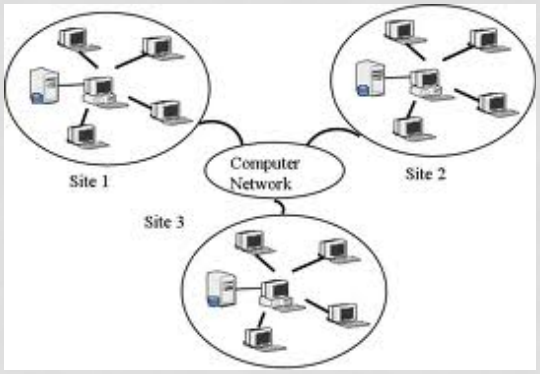


XQuery example

XSLT example

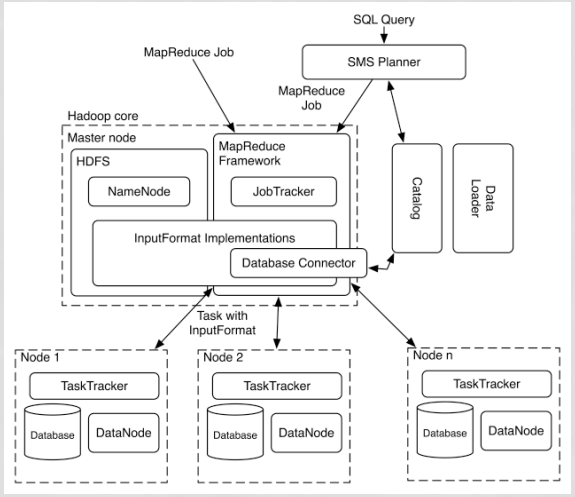
3-DISTRIBUTED AND PARALLEL DATABASES

- **Traditional Distributed Databases**
  - Distributed transactions
  - Distributed concurrency control and two-phase commit
  - Distributed query processing



Distributed DB

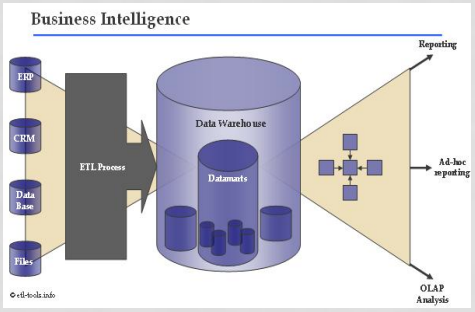
- **Hadoop/MapReduce Infrastructure**
  - New computing paradigm with high scalability, flexibility and fault tolerance
  - Storage paradigm (HDFS)
  - Computing paradigm (Map phase & Reduce phase)



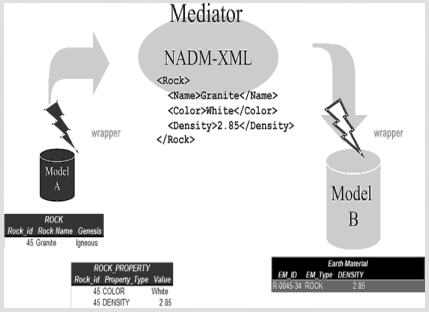
Hadoop Infrastructure

4-INFORMATION INTEGRATION & OLAP

- Data exist in multiple sources (databases or others)
- Information integration is about merging (integrating) the data from all these sources
  - Make all data query-able
  - E.g., Kayak (search engine for hotels/flights) integrates data from many sources
- **Three main architectures**
  - **Federated database**
    - Databases are independent of each other
    - But there a communication link between the individual sources
  - **Data warehousing:**
    - One storage (warehouse) materializing all data (possibly aggregated)
    - Issues of periodic updates
  - **Mediation**
    - Virtual database (with a virtual schema), has no data
    - It routes a query (after transformation) to each source, and then composes the final answer to the individual ones



Data warehouse

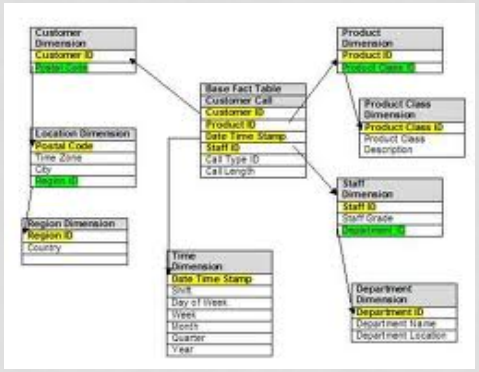


Mediation

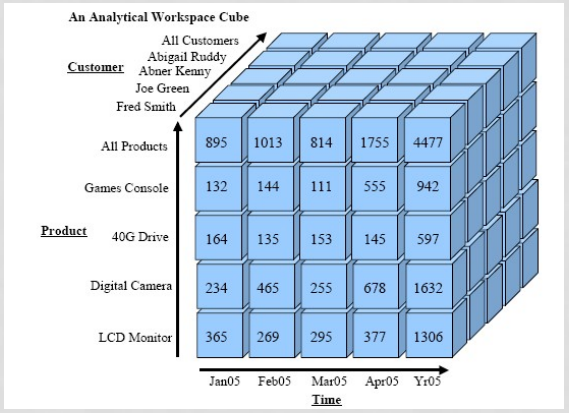
4-INFORMATION INTEGRATION & OLAP

- **OLAP: Online Analytic Processing**
  - Complex queries involving aggregations over one or more dimensions of the data
  - Touch large amount of data for discovering patterns

- **Two important concepts**
  - **Star schema:** one fact table and multiple dimension tables
  - **Data cubes:** data aggregated over different dimensions



Star schema



Data cubes

COURSE LOGISTICS

COURSE MANAGEMENT

- **Web page:** http://web.cs.wpi.edu/~cs561/s12/
- **WPI electronic system**
  - Blackboard pilot: https://blackboard.wpi.edu/
- **Lectures**
  - Tuesday/Thursday: (4:00pm -5:20pm)
  - Location: SL-407
- **Office Hours**
  - Tuesday/Thursday: (2:00pm -3:00pm)
  - Location: My office FL-235

- No required textbook
- Depend on slides + papers + scanned documents that will be posted

- Course content (slides, presentations) will be available on both systems
- Homework submissions, discussions among students, and grading will be within blackboard system

COURSE LOAD

- **Homework (10%)**
  - 4 short homeworks covering the topics given by the instructor
  - Tentative release dates available on the website
- **Presentation (25%)**
  - 2 presentations in the semester ---Select dates
- **Reviews & Participation (15%)**
  - Will talk more about this task
  - Basically, when another student is presenting, you should go over the paper and submit a 1-page review
  - Participate in the class discussion
- **Final exam (15%)**
  - Covering the topics given by the instructor
- **One semester-long project (35%)**

Items	Percentage
Presentations	25%
Reviews & Participation	15%
Semester-Long Project	35%
Homeworks	10%
Final Exam	15%

LATE POLICY

- **Homework**
  - One-day late submission is accepted with 10% off the max grade.
  - Two-day late submission is accepted with 20% off the max grade.
  - Beyond that, no late submission is accepted.
- **Reviews**
  - No late submission is accepted.
  - Each student may skip at most two reviews without affecting his/her grade.
- Policy is available on the website (under ***Grading*** tab)

PRESENTATIONS

- Several candidate papers in different areas are available on the website
- **Select your topic of interest + lecture slot**
  - Then discuss with the instructor which paper to cover
- **Paper to be presented should be scheduled at least one week before the presentation**
  - So others can prepare a review
- **First-come-first-served**
  - Empty slots will be assigned by the instructor
- Hints for good presentation are available on the website (under ***Grading*** tab)

EXPECTED SCHEDULE

Week	Day	Topic(s)	Links	Readings	Comments	Presenter
Week 1	01/12/2012	**No Class on Jan12. Our first meeting is on Jan 17.**				
Week 2	01/17/2012	Introduction, Logistics, and Topics Overview				Instructor
	01/19/2012	Object-Oriented Data Model and Querying				Instructor
Week 3	01/24/2012	Object-Oriented Data Model and Querying			Homework 1 is out	Instructor
	01/26/2012	Distributed and Parallel Databases				Instructor
Week 4	01/31/2012	Distributed and Parallel Databases				Instructor
	02/02/2012	Hadoop & MapReduce Infrastructure			Homework 2 is out	Instructor
Week 5	02/07/2012	Paper: Hadoop-related			Student Presentation	
	02/09/2012	Paper: Hadoop-related			Student Presentation	
Week 6	02/14/2012	Scientific Data Management				Instructor
	02/16/2012	Paper: Scientific Data Management-related			Student Presentation	
Week 7	02/21/2012	Paper: Scientific Data Management-related			Student Presentation	
	02/23/2012	Paper: Keyword Search in Databases-related			Student Presentation	
Week 8	02/28/2012	Paper: Keyword Search in Databases-related			Student Presentation	
	03/01/2012	Active DBs + Authorization				Instructor
Week 9	03/06/2012	Break				
	03/08/2012					
Week 10	03/13/2012	Information Integration + OLAP				Instructor
	03/15/2012	Information Integration + OLAP			Homework 3 is out	Instructor
Week 11	03/20/2012	Paper: Topic of choice			Student Presentation	
	03/22/2012	Paper: Topic of choice			Student Presentation	
Week 12	03/27/2012	Paper: Topic of choice			Student Presentation	
	03/29/2012	Semi-Structured Data Model & Querying				Instructor
Week 13	04/03/2012	Semi-Structured Data Model & Querying			Homework 4 is out	Instructor
	04/05/2012	Paper: Topic of choice			Student Presentation	
Week 14	04/10/2012	Paper: Topic of choice			Student Presentation	
	04/12/2012	Paper: Topic of choice			Student Presentation	
Week 15	04/17/2012	Paper: Topic of choice			Student Presentation	
	04/19/2012	Final Exam				
Week 16	04/24/2012	Project Presentations				
	04/26/2012	Project Presentations				

REVIEWS

- **When a student is presenting a paper, others are reviewing that paper**
  - Reading and understanding the paper
  - Preparing a 1-page review
  - This process will help the discussion in the lecture

- **Structure of good review**
  - **Summary (one paragraph 5-10 lines):** describe briefly the addressed problem and main challenges, and the solution.
  - **Strong Points (2-3 points):** why this work is novel, what is the most interesting idea behind the solution, does the paper have enough evaluation and performance measures.
  - **Weak Points (2-3 points):** what do you think have not been addressed adequately, possible weaknesses, assumptions that are not practical, or extensions you think are good.

PROJECT

- Teams of 2 (or 3)
- Several candidate projects to select from (or come up with new ideas)
- **Platform to work on:**
  - PostgreSQL, or
  - Hadoop
- Work closely with instructor for continues feedback and directions
- Study and comparison between different techniques or exploring new ideas
- **By next Thursday (Jan. 26) groups should be formed and the project is selected**