Study Guide: Data Models

DSC 208R - Data Management for Analytics

Module Overview

This guide covers fundamental data models with focus on relational and DataFrame paradigms. Key topics include structural components, constraints, and SQL operations.

1 DataFrame Model

Core Concepts

Historical Development:

- 1992: Originated in S language (Bell Labs)
- 2000: Adopted by R
- 2009: Pandas implementation in Python

Key Features:

- Hybrid operations: Relational + Linear Algebra + Spreadsheet
- Labeled axes (rows & columns)
- Heterogeneous data types per column

Comparative Analysis

Aspect	vs. Relational	vs. Matrices
Schema	Lazily-induced	N/A
Structure	Named/ordered rows & columns	Numeric indices
Data Types	Heterogeneous columns	Homogeneous elements
Operations	Filter/Join + Transpose + Pivot	Pure linear algebra

2 Relational Model

Structural Components

```
CREATE TABLE Students (
    sid     CHAR(20) PRIMARY KEY,
    name    CHAR(30),
    age     INTEGER,
    gpa     REAL
);
```

Core Elements:

- **Relation**: Table with attributes (columns) and tuples (rows)
- **Schema**: Structural metadata (name:type pairs)
- **Instance**: Current dataset conforming to schema

Constraints

Domain Constraints:

- Enforce data types (INT, CHAR, DATE)

Key Constraints:

- Candidate Key: Minimal unique identifier
- Primary Key: Chosen main identifier
- Super Key: Superset containing candidate key

Referential Integrity:

```
CREATE TABLE Enrolled (
    sid CHAR(20) REFERENCES Students(sid),
    ...
);
```

3 SQL Fundamentals

Essential Operations

Operation	SQL Example
Create Table	`CREATE TABLE Students (...);`
Insert Data	`INSERT INTO Students VALUES (...);`
Delete	`DELETE FROM Students WHERE age > 30;`
Update	`UPDATE Students SET gpa = 3.5 WHERE ...;`
Query	`SELECT name, gpa FROM Students WHERE ...;`
Alter	`ALTER TABLE Students ADD email VARCHAR;`

First Normal Form (1NF)

Requirement: Atomic values, no nested/repeating groups
Violation Example:

```
CREATE TABLE BadDesign (
    sid INT,
    courses_enrolled ARRAY   -- Invalid
);
```

1NF Solution:

```
CREATE TABLE Enrolled (
    sid INT REFERENCES Students,
    cid CHAR(10),
    grade REAL
);
```

4 Key Takeaways

Essential Concepts

1. **DataFrame Model** bridges relational and numerical computing
2. **Relational Model** requires explicit schema + constraints
3. **1NF** ensures atomic values through flat table structures
4. **SQL** enables declarative data definition and manipulation