**Big data notes**

**Introduction to Data model**

**What is a Data Model​?**

Conceptual representation of the data structures that are required by a database.

The data structures include​

* the data objects​
* the associations between data objects​
* the rules which govern operations on the objects

**Different Data Models​**

* Conceptual: describes WHAT the system contains​
* Logical: describes HOW the system will be implemented, regardless​ of the DBMS​
* Physical: describes HOW the system will be implemented using a ​ specific DBMS

Elements of Data Models​

* Entity - A real world thing or an interaction between 2 or more real world things​
* Attribute - The atomic pieces of information that we need to know about entities​
* Relationship  - How entities depend on each other in terms of why the entities depend on each other and what that relationship is

Conceptual Data Model​

* The focus is to represent data as a user will see it in the "real world" ​
* The main aim of this model is to establish the entities, their attributes, and their relationships.

Logical Data Model​

* Logical data models add further information to the conceptual model elements.​
* Defines the structure of the data elements and set the relationships between them

Physical Data Model​

* It describes the database specific implementation of the data model.​
* It offers an abstraction of the database and helps generate schema.

Advantages of Data Models​

* ​To make sure that the data objects offered by the functional team are represented accurately.​
* Should be detailed enough to be used for building the physical database.​
* Information in the data model can be used for defining the relationship between tables, primary and foreign keys, and stored procedures.​
* It helps business to communicate within and across organizations.​
* It helps to recognize correct sources of data to populate the model.

Disadvantages

* To develop a Data model, one should know the physical characteristics of the data.​
* It requires knowledge of the hidden truth related to the system.​
* Even smaller change made in structure require modification in the entire application.​
* There is no set data manipulation language in DBMS.

Common data modelling notations​

1. Barker’s Notation​
2. Chen Notation​

**In Chen Notation, we distinguish types of entities:**

Entity - rectangle

Weak Entity

Associative Entity

**Attributes – Oval**

Key Attribute (underscored)

Partial Key Attribute (dashed underscored)

**Multivalued Attribute (ex: hobby)** -  Multivalued attribute is depicted by a dual oval

**Derived Attribute (dashed Oval)**

Some attributes can be further subdivided into smaller parts and are called**composite attributes**

**Strong relationship - single rhombus**

**Weak (identifying) relationship -**A relationship where Child entity is existence-dependent on parent, and PK of Child Entity contains PK component of Parent Entity​. This relationship is represented by a double rhombus

**Cardinality**

**Participation Constraints**

1. IDEF1X Notation​

**IDEF1X** (Integration DEFinition for Information Modeling) is a method for designing relational databases

1. Arrow Notation​
2. UML Notation​

**Generalization** is represented by an empty arrow​

**Aggregation** is shown as binary association with a hollow diamond as a symbol of the aggregation at the end of the association line​

**Composition** is presented as binary association with a black diamond as a symbol of the composition at the end of the association line

1. Crow’s Foot Notation​

​Also known as IE notation

* An entity is a representation of a class of object ​
* It can be a person, place, thing, etc​
* Entities usually have attributes that describe them​
* In crow’s foot notation, an entity is represented by a rectangle, with its name on the top​
* The name is singular (entity) rather than plural (entities)

**How to Model Data:​**

1. Identify entity types​
2. Identify attributes​
3. Apply naming conventions​
4. Identify relationships​
5. Apply data model patterns​
6. Assign keys​
7. Normalize to reduce data redundancy

**Elementary data analysis**

**Apache Mahaout**

**Apache Hive**

**AWS quicksight**

**Corelation Pattern**

**Bayesian staistics**