


(Big) Data Engineering In Depth

From Beginner to Professional

Mostafa Alaa Mohamed

Senior Big Data Engineer

 MoustafaAlaa  Moustafa Alaa  @Moustafa_alaa22

 mustafa.alaa.mohamed@gmail.com

¹Big Data & Analytics Department, Epam Systems

The Definitive Guide to Big Data Engineering Tasks

Videos classification

Watching Method / Audience	Computer	Mobile/Tablet	Just listening
Developer	●		
DevOps	●		
Business	●		

Table: Video classification

The green circle ● means short video.
The blue circle ● means medium video.
The red circle ● means long video

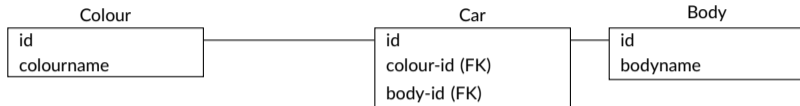
Dimensions Types: Junk Dimension (Garbage Dimension)

Junk Dimension

- It used to reduce the number of dimensions (low-cardinality columns) in the dimensional model and reduce the number of columns in the fact table. It is a collection of random transnational codes, flags, or text attributes.
- It optimizes space as fact tables should not include low-cardinality or text fields. It mainly includes measures, foreign keys, and degenerate dimension keys.

Junk Dimension

Design without junk DIM



Design with junk DIM



Junk Dimension Table Size

- We must split the Junk dimension into more dimensions in case the size grows by the time.
- It is easy to calculate the expected number of rows as it is the total number of combinations between the low-cardinality attributes;
✍️ ➡️ 3 columns each have 3 values total = $3 * 3 = 9$.