


(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: Multi-valued dimensions (Many-To-Many Dimension)

Multi-valued dimensions

- When the relationships between the dimension member and the fact are many to many which means the dimension members are lower granularity than the facts.

Example

Multi-valued dimensions

- When the relationships between the dimension member and the fact are many to many which means the dimension members are lower granularity than the facts.
- Fact table should contains one-to-one relationship with the dimension. So, we introduce the ***Bridge table*** when we need to related multiple dimensions values with one record.

Example

Multi-valued dimensions

- When the relationships between the dimension member and the fact are many to many which means the dimension members are lower granularity than the facts.
- Fact table should contains one-to-one relationship with the dimension. So, we introduce the ***Bridge table*** when we need to related multiple dimensions values with one record.

Example

- Patients can have multiple diagnoses.

Multi-valued dimensions

- When the relationships between the dimension member and the fact are many to many which means the dimension members are lower granularity than the facts.
- Fact table should contains one-to-one relationship with the dimension. So, we introduce the **Bridge table** when we need to related multiple dimensions values with one record.

Example

- Patients can have multiple diagnoses.
- Students can have multiple majors.

Multi-valued dimensions

- When the relationships between the dimension member and the fact are many to many which means the dimension members are lower granularity than the facts.
- Fact table should contains one-to-one relationship with the dimension. So, we introduce the **Bridge table** when we need to related multiple dimensions values with one record.

Example

- Patients can have multiple diagnoses.
- Students can have multiple majors.
- customers can have multiple account.

Multi-valued dimensions

- When the relationships between the dimension member and the fact are many to many which means the dimension members are lower granularity than the facts.
- Fact table should contains one-to-one relationship with the dimension. So, we introduce the **Bridge table** when we need to related multiple dimensions values with one record.

Example

- Patients can have multiple diagnoses.
- Students can have multiple majors.
- customers can have multiple account.
- Authors can have multiple publications.

Multi-valued dimensions

Example (Sales of Articles)

- Assume we need to report the sales of article and we have some articles has more than one author.

ID	Name	Email	Bio
123	Moustafa	abc@gability.com	S-Engineer
234	Ahmed	def@gability.com	L-Engineer
345	Amr	geh@gability.com	S-Manager

ID	Title	Journal	Price
11	50	IEEE	110.0
22	55	ACM	130.0

Table: author and articles sample data.

Multi-valued dimensions

Example (Sales of Articles)

- Assume we need to report the sales of article and we have some articles has more than one author.
- Each author has weighting factor for each article.

ID	Name	Email	Bio
123	Moustafa	abc@gability.com	S-Engineer
234	Ahmed	def@gability.com	L-Engineer
345	Amr	geh@gability.com	S-Manager

ID	Title	Journal	Price
11	50	IEEE	110.0
22	55	ACM	130.0

Table: author and articles sample data.

Multi-valued dimensions

Example (Sales of Articles)

- Assume we need to report the sales of article and we have some articles has more than one author.
- Each author has weighting factor for each article.
- According to the report we need to check each author and associate with the articles they have authored. How can we model this case?

ID	Name	Email	Bio
123	Moustafa	abc@gability.com	S-Engineer
234	Ahmed	def@gability.com	L-Engineer
345	Amr	geh@gability.com	S-Manager

ID	Title	Journal	Price
11	50	IEEE	110.0
22	55	ACM	130.0

Table: author and articles sample data.

Multi-valued dimensions

Example (Sales of Articles)

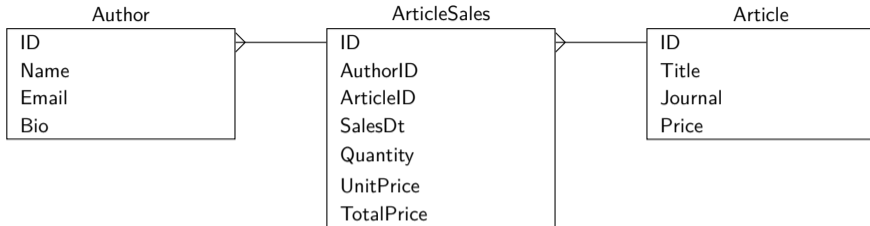
- Assume we need to report the sales of article and we have some articles has more than one author.
- Each author has weighting factor for each article.
- According to the report we need to check each author and associate with the articles they have authored. How can we model this case?
- Assume the first article has only one author *Moustafa*, and the second article has two authors *Ahmed & Amr*.

ID	Name	Email	Bio
123	Moustafa	abc@gability.com	S-Engineer
234	Ahmed	def@gability.com	L-Engineer
345	Amr	geh@gability.com	S-Manager

ID	Title	Journal	Price
11	50	IEEE	110.0
22	55	ACM	130.0

Table: author and articles sample data.

Multi-valued dimensions (Implementation-1)



Multi-valued dimensions (Implementation-1)



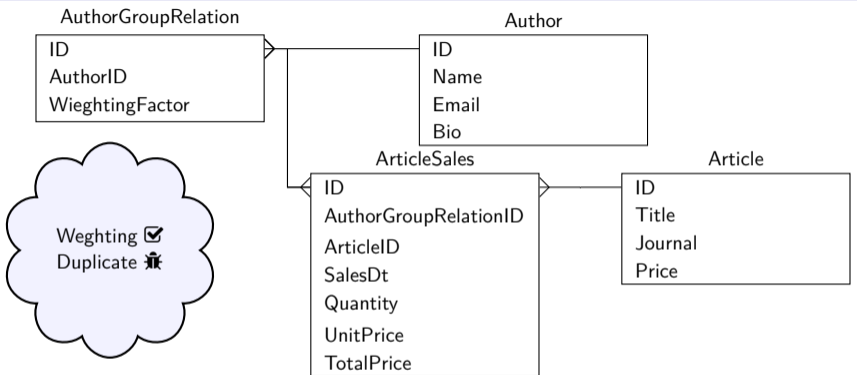
ID	AuthorID	ArticleID	SalesDt	Quantity	UnitPrice	TotalOrder
1	123	11	20200303	3	10	30
2	234	22	20200304	1	20	20
3	345	22	20200304	1	20	20

Table: Output of wrong implementation of ArticleSales

What are the problems in this implementation?

- We can't get the weighting factor for each author.
- Duplicated rows in sales.

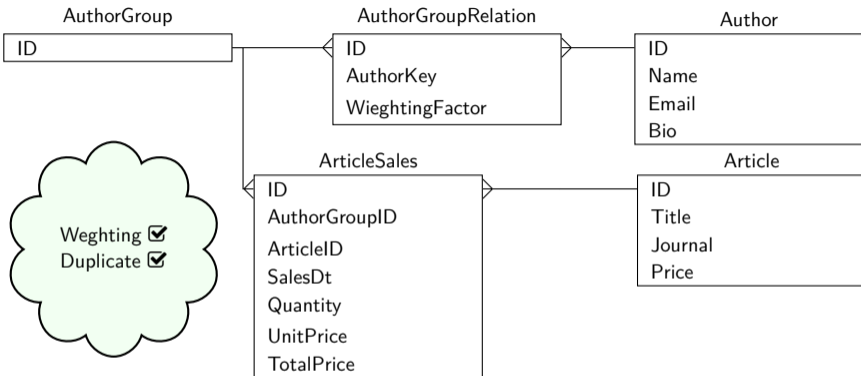
Multi-valued dimensions (Implementation-2)



ID	AuthorGroupRelationID	ArticleID	SalesDt	Quantity	UnitPrice	TotalOrder
1	321	11	20200303	3	10	30
2	432	22	20200304	1	20	20
3	432	22	20200304	1	20	20

Table: Output of wrong implementation of ArticleSales

Multi-valued dimensions (Final Implementation)



ID	AuthorGroupID	ArticleID	SalesDt	Quantity	UnitPrice	TotalOrder
1	321	11	20200303	3	10	30
2	432	22	20200304	1	20	20

Table: Expected output of ArticleSales

Example Reference

- Example in this video taken from this link
<https://www.nuwavesolutions.com/bridge-tables/>