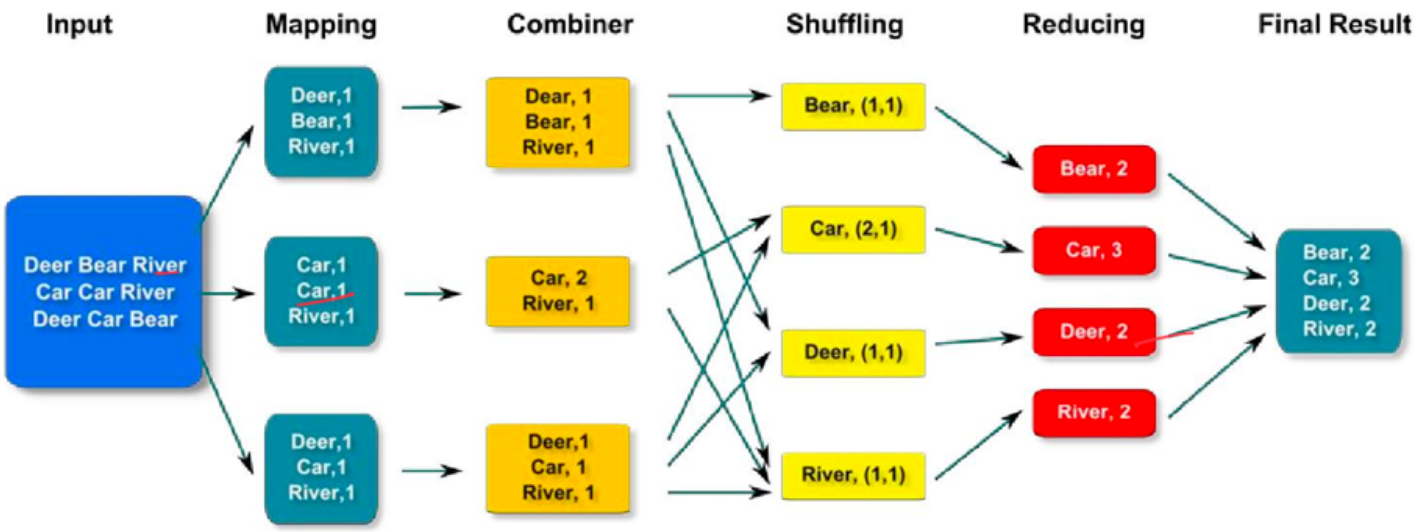
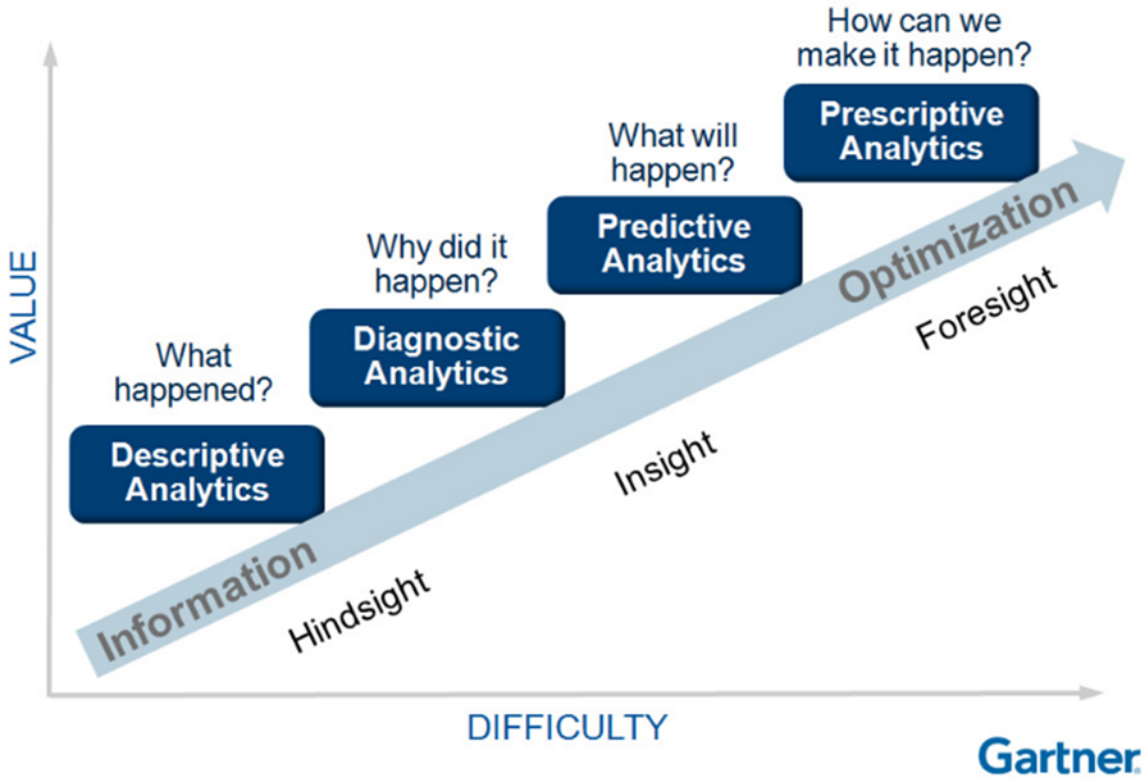
Agenda

Map reduce optimisations
What is OLTP?
Why it is not suitable for Analytics
Introduction to Data warehousing
Why is data warehouse a better option?
How to design a Data Warehouse
Dimensional Modeling Approach

Combiner - Local Reduce

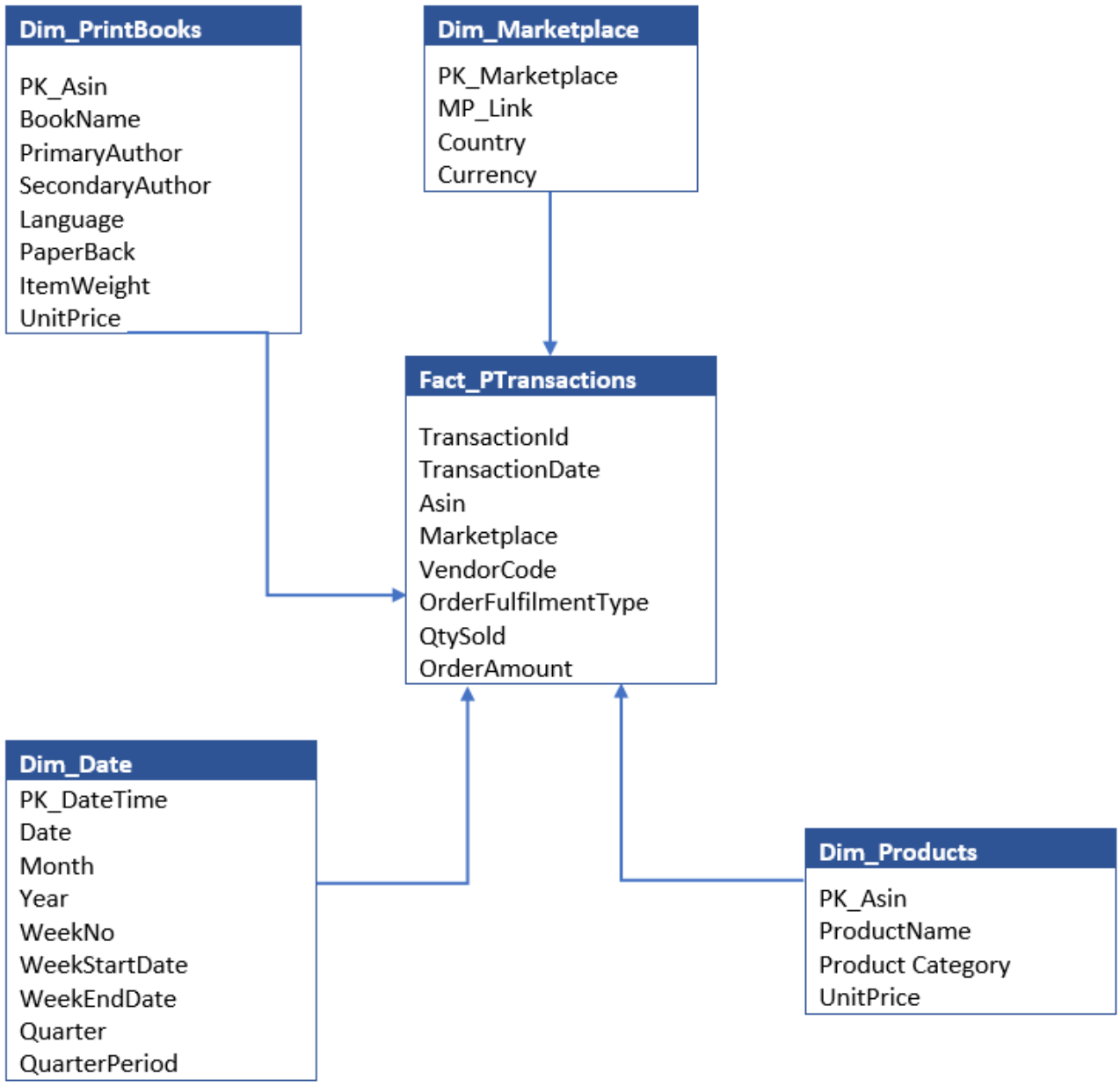


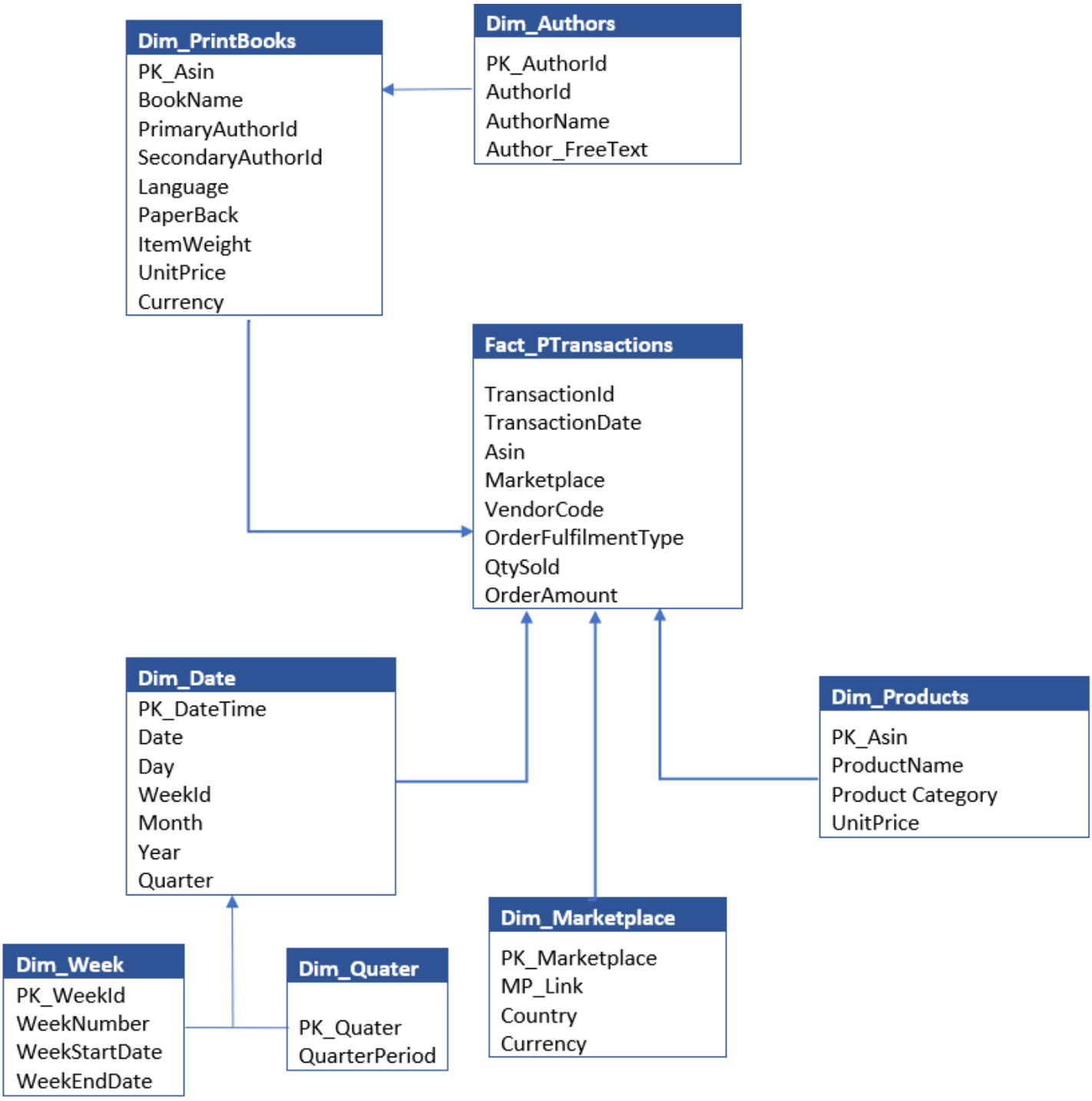
Analytic Value Escalator

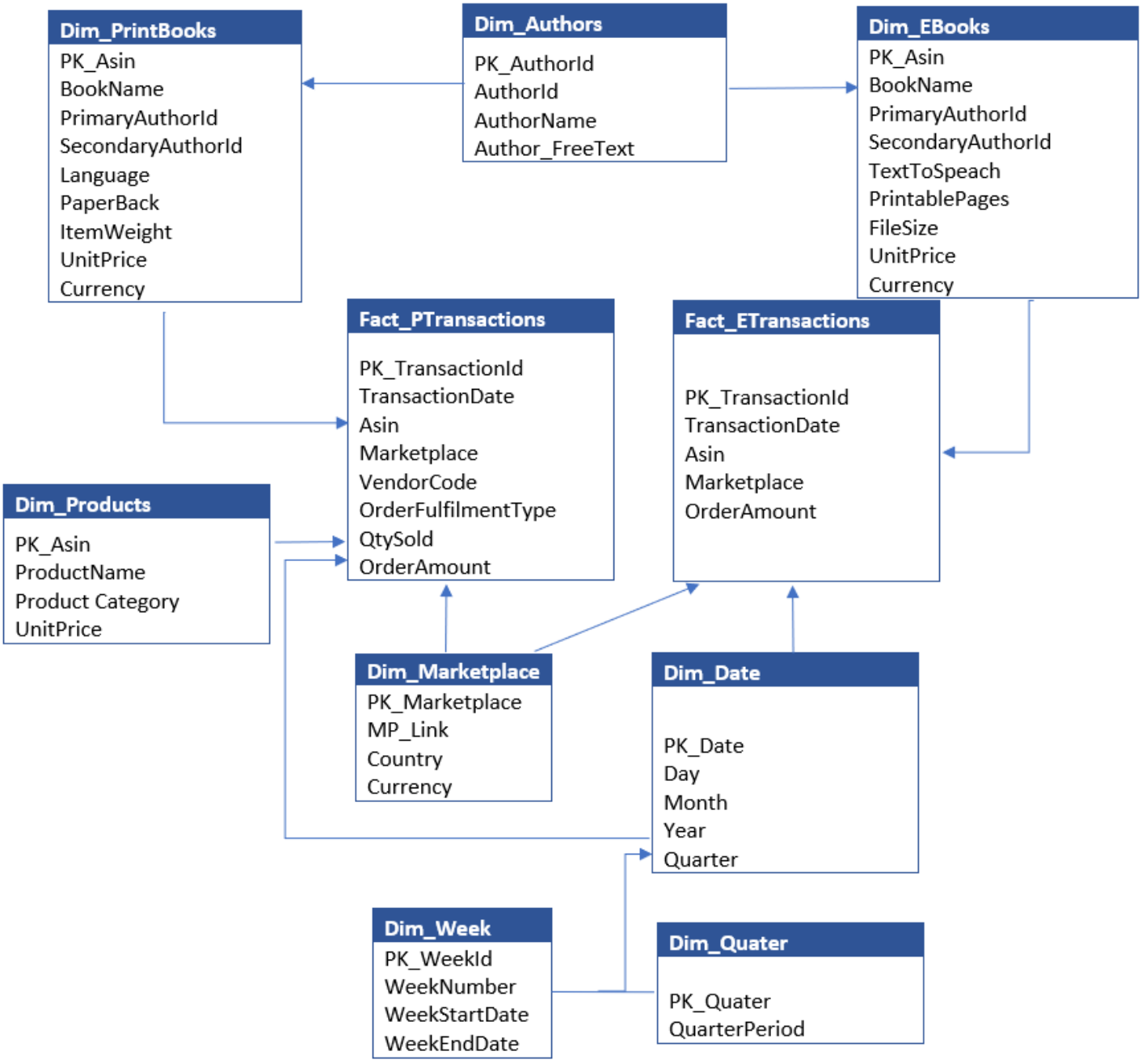


OrderId	CustomerId	InvoiceNo	Invoice Date	ShippingAddress	PaymentMethod	SoldBy	BillingAddress	PlaceOfSupply	PlaceOfDelivery	ASIN	Item.No	Description	UnitPrice	Qty	Net Amount	Tax Amount	Total Amount	Shipping
7146	101	0990	23 August 2022	Bangalore	COD	CT	Bangalore	Chennai	Bangalore	B0963RYQSW	1	Adapter	100	1	100	18	118	0

OrderId	CustomerId	InvoiceNo	Invoice Date	ShippingAddress	PaymentMethod	SoldBy	BillingAddress	PlaceOfSupply	PlaceOfDelivery	ASIN	Item.No	Description	UnitPrice	Qty	Net Amount	Tax Amount	Total Amount	Shipping
7146	101	0990	23 August 2022	sjkgfjsvfjsvn	COD	CT	sjkgfjsvfjsvn	Chennai	sjkgfjsvfjsvn	B0963RYQSW	1	Adapter	100	1	100	18	118	0

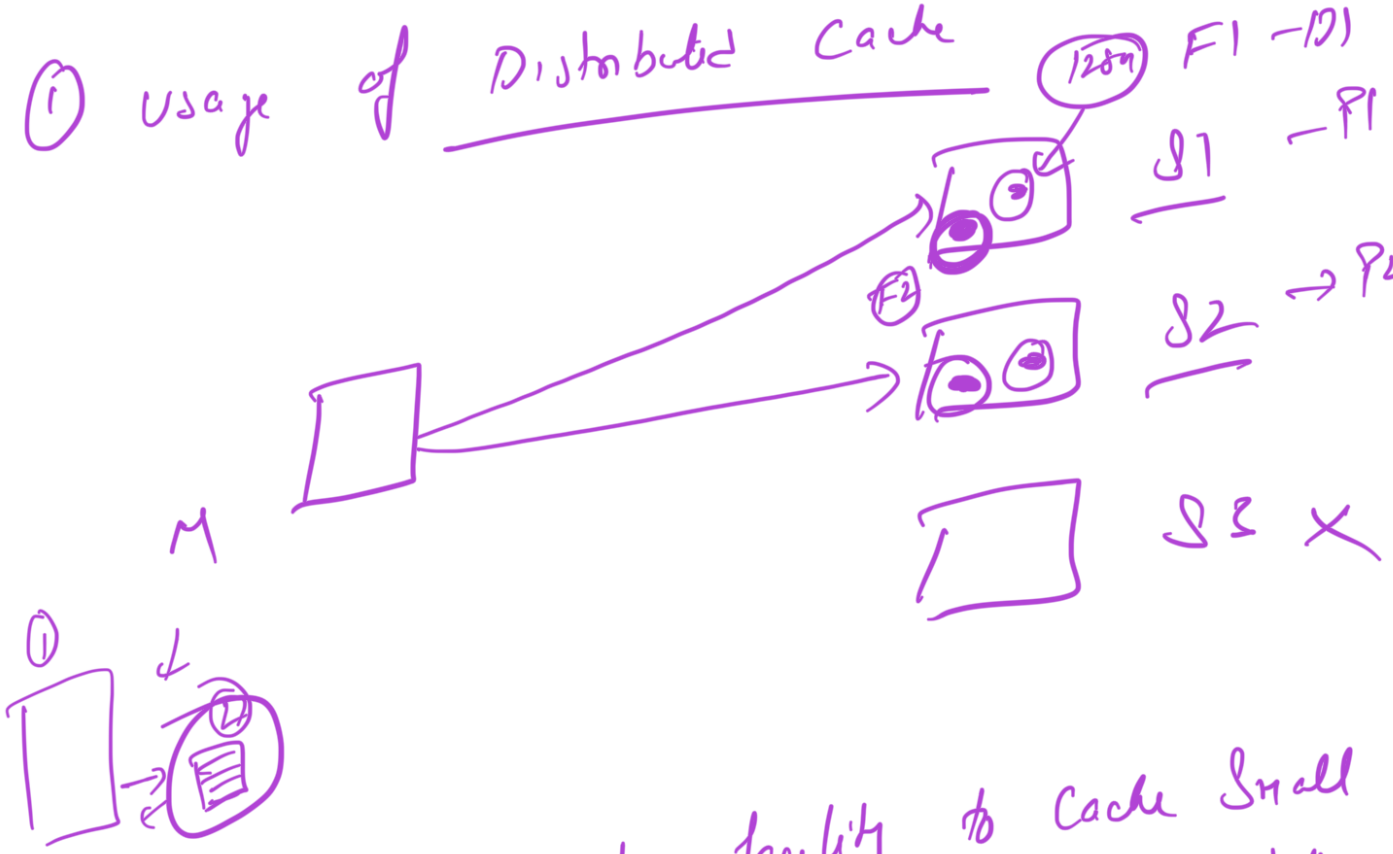



Dim_PrintBooks
PK_Asin
BookName
PrimaryAuthorId
SecondaryAuthorId
Language
PaperBack
ItemWeight
UnitPrice
Currency
Dim_Authors
PK_AuthorId
AuthorId
AuthorName
Author_FreeText
Fact_PTransactions
TransactionId
TransactionDate
Asin
Marketplace
VendorCode
OrderFulfilmentType
QtySold
OrderAmount
Dim_Date
PK_DateTime
Date
Day
WeekId
Month
Year
Quarter
Dim_Products
PK_Asin
ProductName
Product Category
UnitPrice
Dim_Week
PK_WeekId
WeekNumber
WeekStartDate
WeekEndDate
Dim_Quater
PK_Quater
QuarterPeriod
Dim_Marketplace
PK_Marketplace
MP_Link
Country
Currency


Dim_PrintBooks
PK_Asin
BookName
PrimaryAuthorId
SecondaryAuthorId
Language
PaperBack
ItemWeight
UnitPrice
Currency
Dim_Authors
PK_AuthorId
AuthorId
AuthorName
Author_FreeText
Dim_EBooks
PK_Asin
BookName
PrimaryAuthorId
SecondaryAuthorId
TextToSpeach
PrintablePages
FileSize
UnitPrice
Currency
Fact_PTransactions
PK_TransactionId
TransactionDate
Asin
Marketplace
VendorCode
OrderFulfilmentType
QtySold
OrderAmount
Fact_ETransactions
PK_TransactionId
TransactionDate
Asin
Marketplace
OrderAmount
Dim_Products
PK_Asin
ProductName
Product Category
UnitPrice
Dim_Marketplace
PK_Marketplace
MP_Link
Country
Currency
Dim_Date
PK_Date
Day
Month
Year
Quarter
Dim_Week
PK_WeekId
WeekNumber
WeekStartDate
WeekEndDate
Dim_Quater
PK_Quater
QuarterPeriod

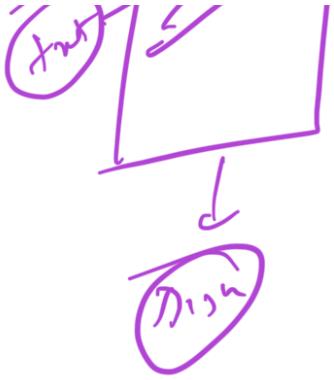
Map Reduce optimizations

① Usage of Distributed Cache



⇒ MR provides facility to cache small to moderate read only files like text, zip, jar.

⇒ Each Datanode gets a copy of the file which is sent through DC. When Jobs gets finished, files are deleted from DN.



5. Take the help from Combiners

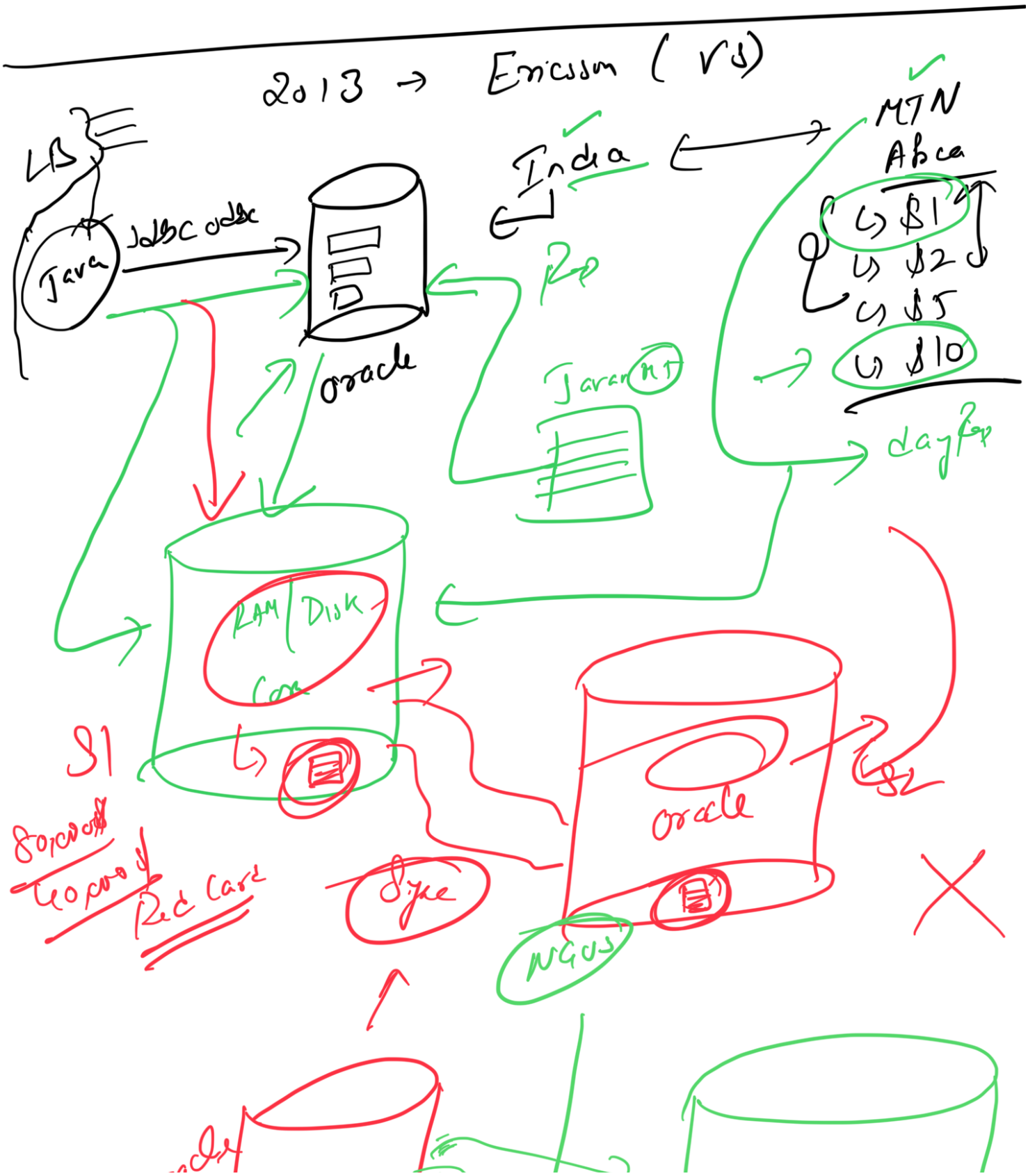
6. mapred.child.java.opts ← Tune
↳ try to avoid swapping
OS → Page Swap

OS
OOM
12GB
↳ 4GB
↳ 8GB → OS

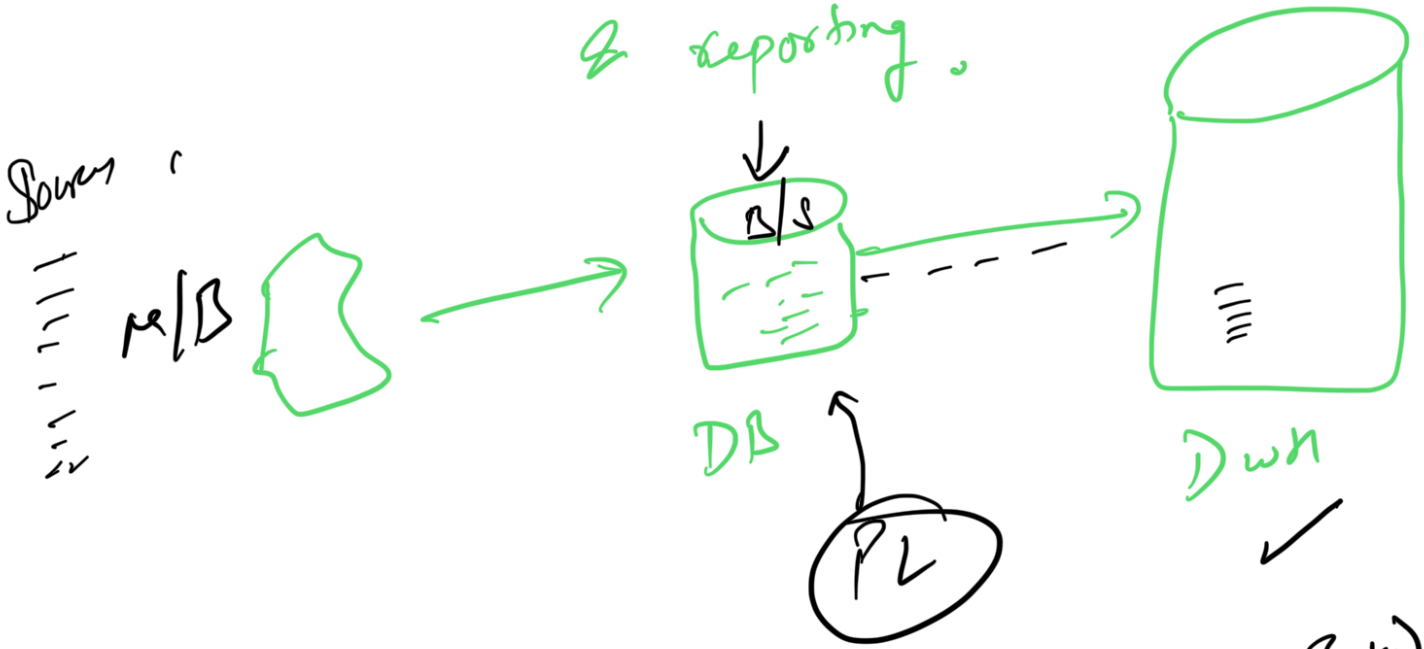
Data Modelling
↳ SQL
↳ Datawarehousing

Amazon Books
↳ physical
↳ e-books

→ what top 10 books sold
→ what other products were frequently bought with books?
→ who are the top authors based on book sales?



DWH :- is a storage unit environment to support BA activities like analysis & reporting.



Design a DWH (Amazon Books)

↳ Books/Books related → Author
↳ Sales → Marketplace
↳ Corporate

↳ Vendor

Dimensional Modeling

↳ is a technique which include a set of methods, practices and concepts to design and store data in a DWH for storage optimization & boost performance.

→ Fact and Dimension

Fact

↳ Facts are measurements / Metrics

Numbers

Books → qty sold
orderAmount
Price
Discount
total stock

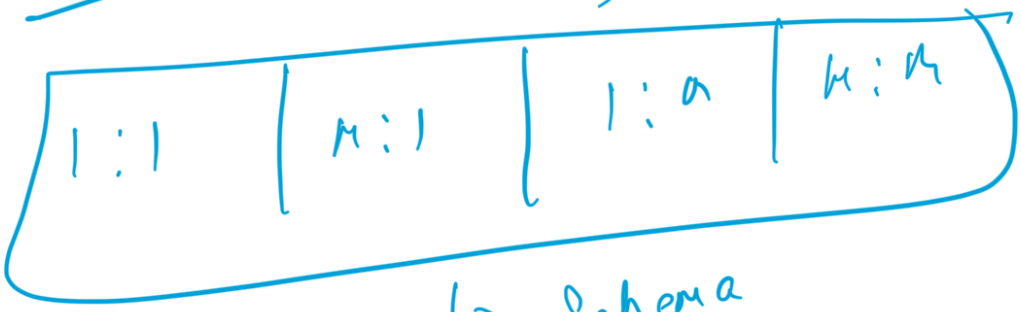
Dimension → provides the context surrounding a business process event, in very simple terms :- when

Who
What
Where
Fact

→ Who – Author, Publishers, Customer
→ Where – location
→ What – Book name

✓ Fact table :- quantitative info

✓ Dimension table :- descriptive info



Sql

↳ Schema

(Sql) (1) Star Schema
(Sql) (2) Snowflake schema
(Sql) (3) Galaxy schema

→ Star Schema :- type of schema that utilizes facts & Dimensions to create a simple yet efficient dimensional

Model.

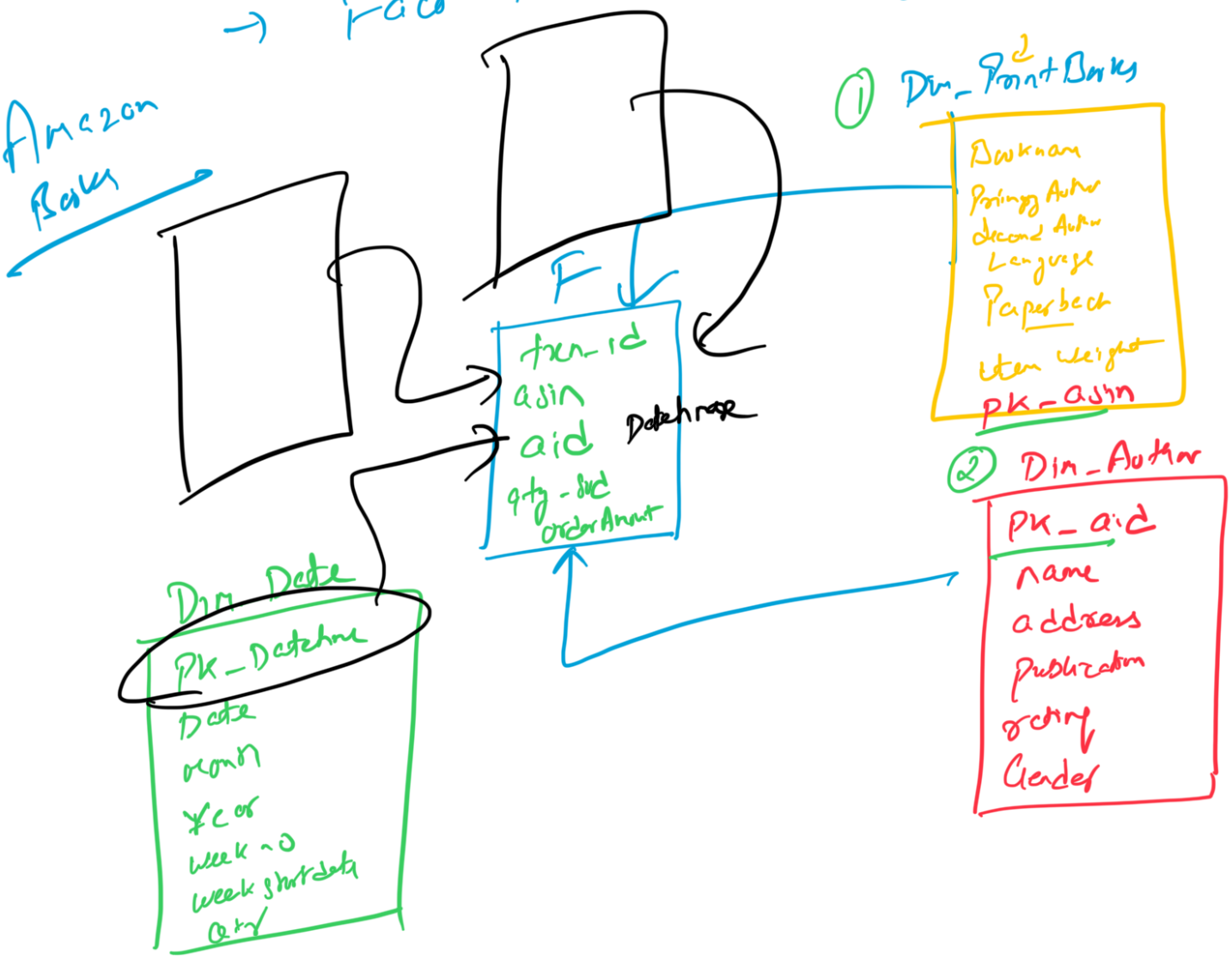
→ Every dimension in a star Schema is represented with only 1 dimension table.

→ Dim. table should contain the set of attributes

→ Dim table is joined to fact table using a FK.

→ Dim table are not joined to each other.

→ Fact table Contain key & measure.



DC	DWH	VDWH
	→ Physical Disk → on-prem	→ Virtual Disk → cloud