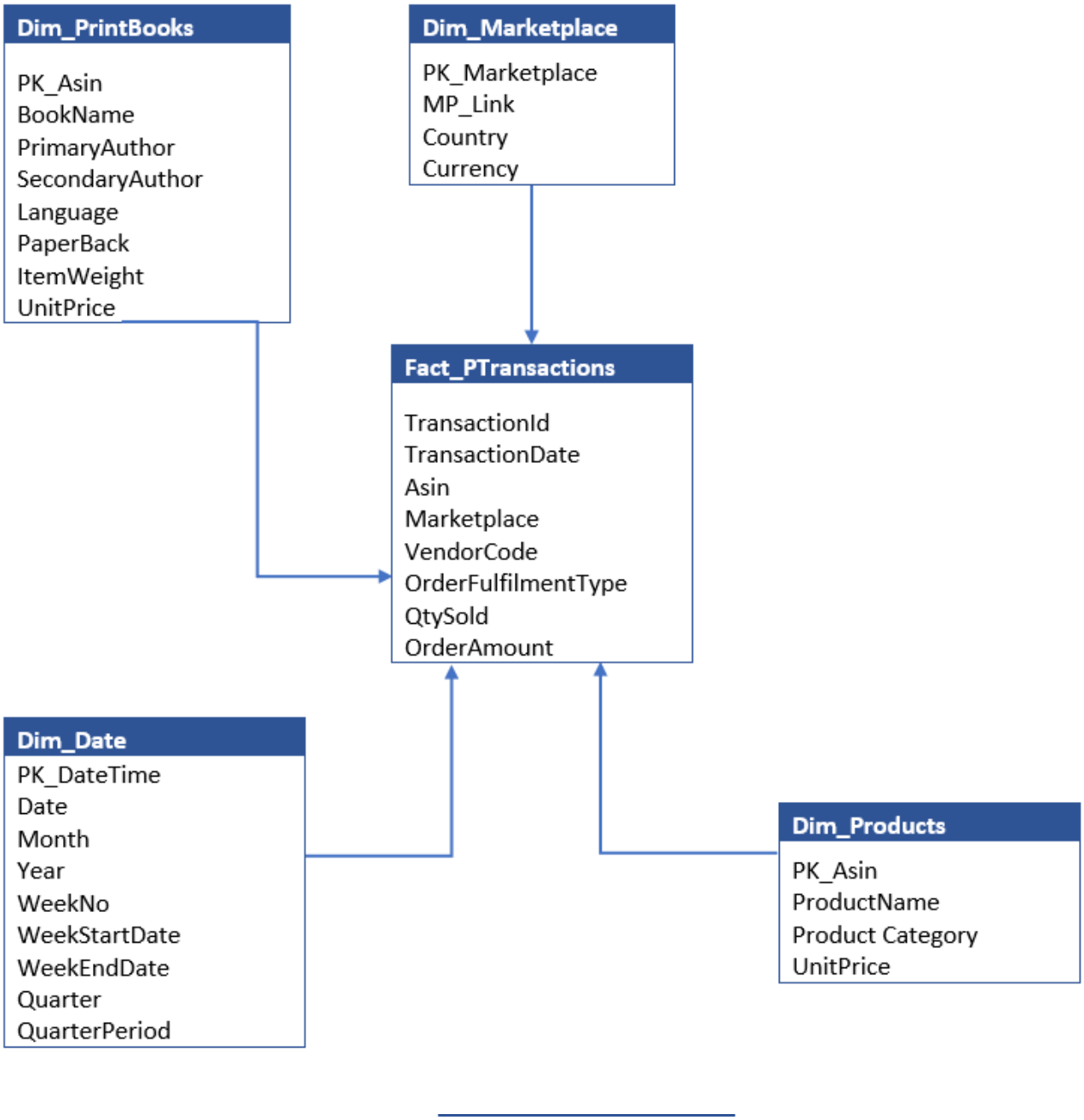
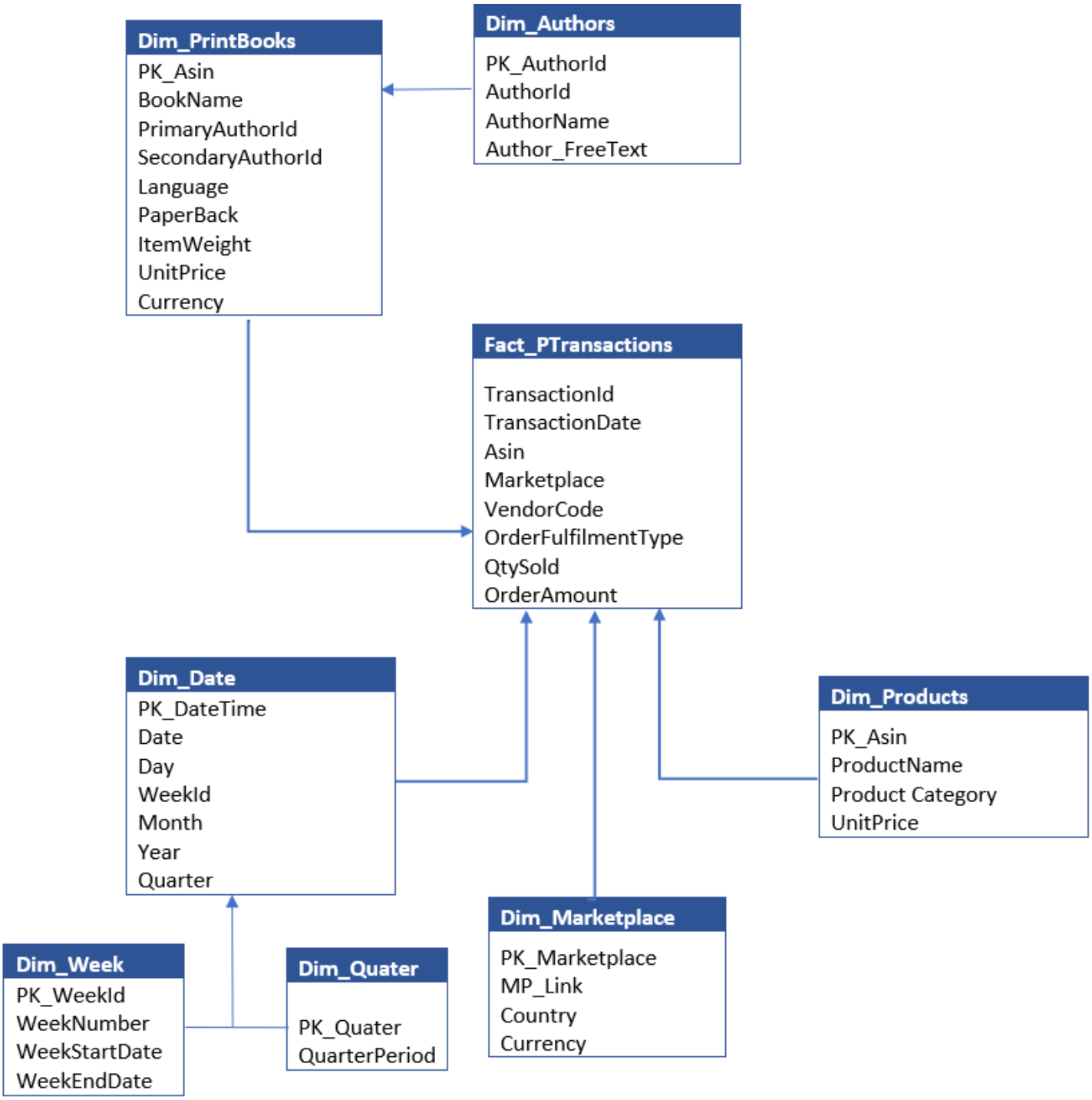
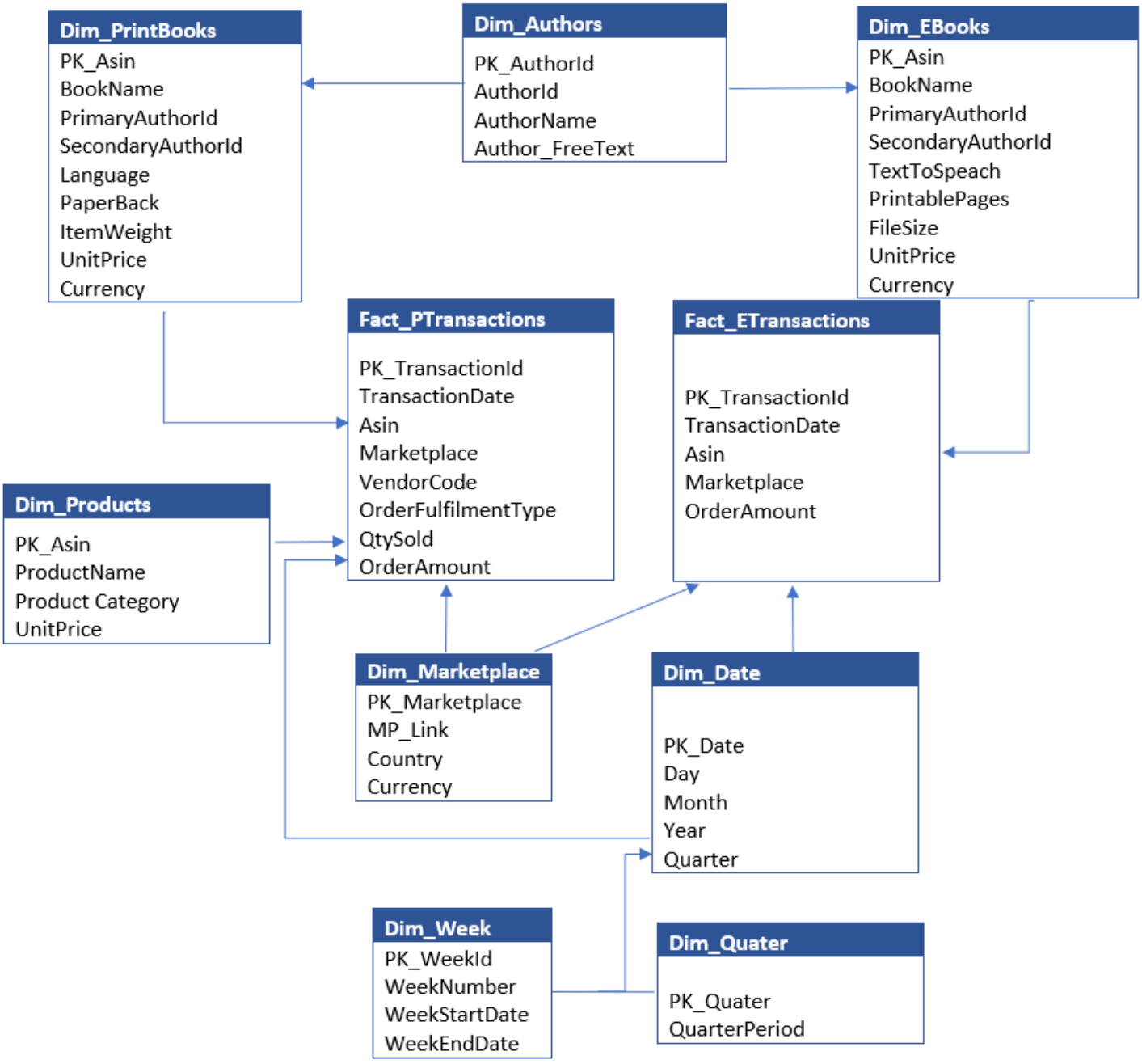
Agenda

Dimensional Modeling Approach

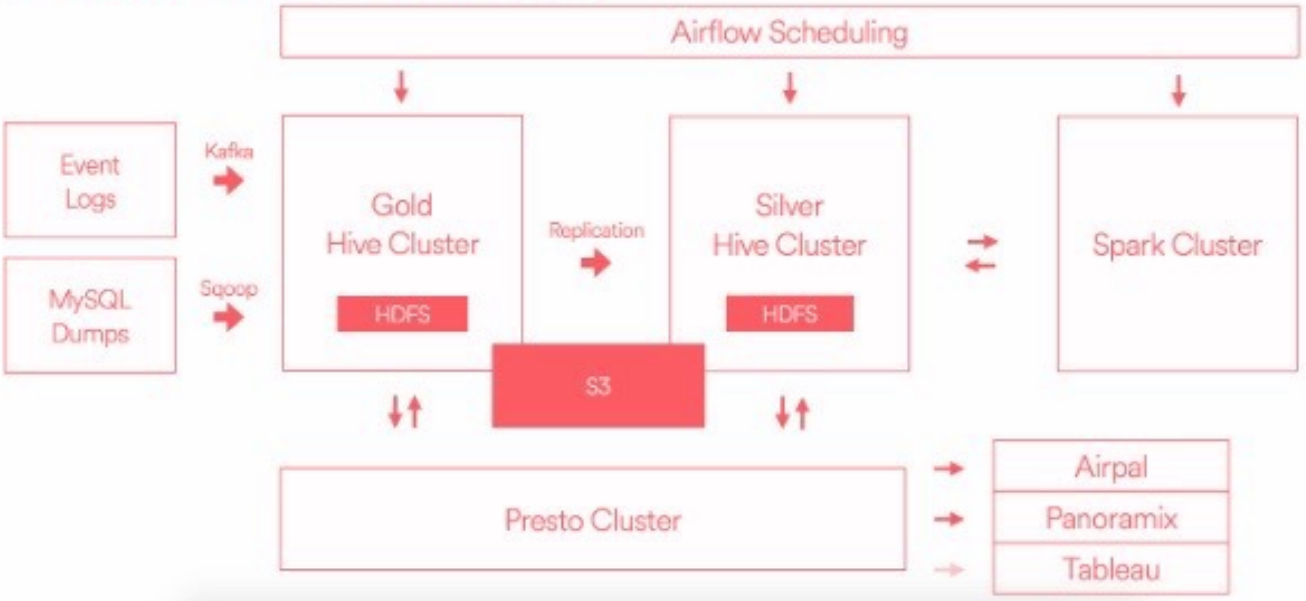
- Problem statement
- Existing Solution
- What are data warehouses and what is OLAP?
- How does Apache hive act as a data warehouse?
- What makes hive as Hive? (Architecture)
- How to interact with Hive?
- What are different HQL commands that are used commonly?
- What are the optimization techniques in Hive ?
- Use cases of hive/ where to use hive/why and why not


Dim_PrintBooks
PK_Asin
BookName
PrimaryAuthor
SecondaryAuthor
Language
PaperBack
ItemWeight
UnitPrice
Dim_Marketplace
PK_Marketplace
MP_Link
Country
Currency
Fact_PTransactions
TransactionId
TransactionDate
Asin
Marketplace
VendorCode
OrderFulfilmentType
QtySold
OrderAmount
Dim_Date
PK_DateTime
Date
Month
Year
WeekNo
WeekStartDate
WeekEndDate
Quarter
QuarterPeriod
Dim_Products
PK_Asin
ProductName
Product Category
UnitPrice

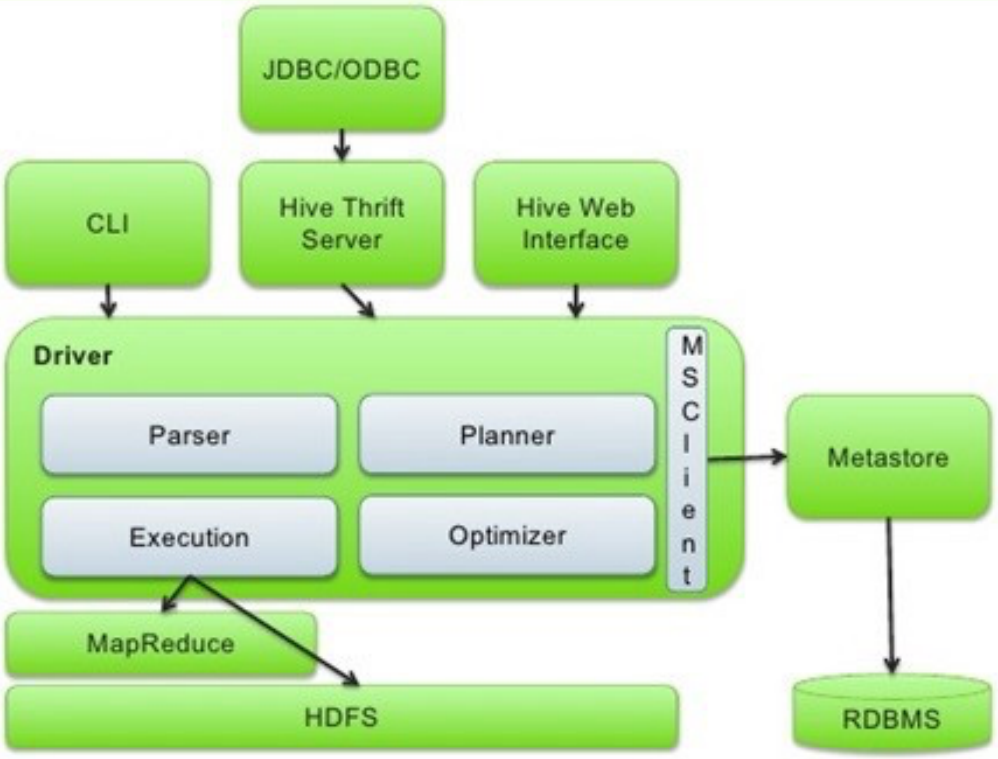

Dim_PrintBooks
PK_Asin
BookName
PrimaryAuthorId
SecondaryAuthorId
Language
PaperBack
ItemWeight
UnitPrice
Currency
Dim_Authors
PK_AuthorId
AuthorId
AuthorName
Author_FreeText
Fact_PTransactions
TransactionId
TransactionDate
Asin
Marketplace
VendorCode
OrderFulfilmentType
QtySold
OrderAmount
Dim_Date
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

AIRBNB DATA INFRA



Apache Hive Architecture



Receive SQL query

1. Parse HiveQL
2. Make optimizations
3. Plan execution
4. Submit job(s) to cluster
5. Monitor progress
6. Process data in MapReduce or Apache Spark
7. Store the data in HDFS

Star Schema

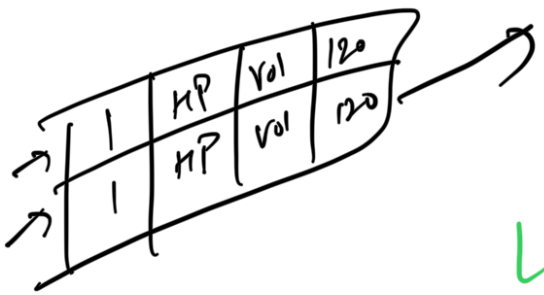
Advantages

↳ Star schema is easy to understand and provides optimal disk usage.

↳ Schema is widely supported by BI Tools.

Disadvantages :-

↳ Redundancy ⇒ Data is denormalized, leading to duplication and increased storage requirements.



↳ Limited flexibility ⇒ less flexible when dealing with complex queries.

↳ Performance Issues ⇒ Large dimension table can slow down performance.

↳ Maintenance challenges :- updates/changes in this schema can be more complex & error prone.

↳ Migration challenges :-

(2) Snowflake Schema:- extended form of star schema, by normalising dimension table.

↳ advantages:- (a) Low Data redundancy
(b) Consumes less storage

Disadvantages:- (a) Complexity
(b) Query Performance is lower
(c) Designing is Complex than star schema.

(3) Galaxy | Fact Constellation Schema

↳ Compound schema which contains more than 1 fact and normalised Dimension.

→ Advantages
↳ Reusability
↳ Supports Complex Business Processing
↳ Scalability

Snowflake

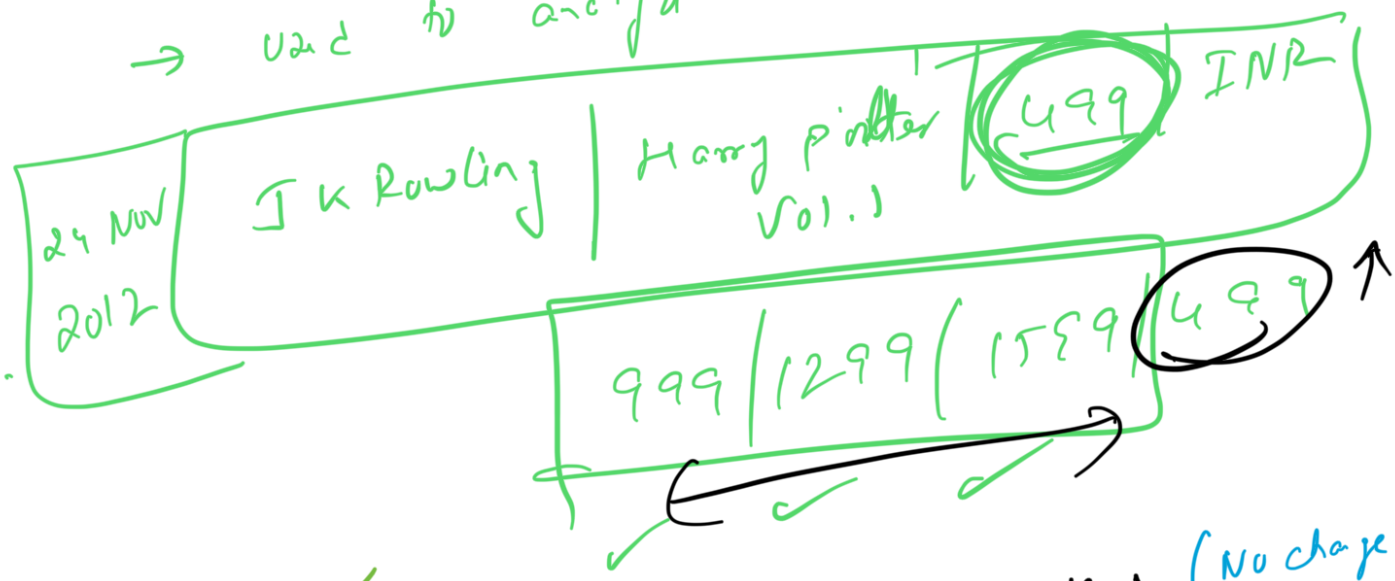
Complexity

→ Disadvantages :- (1) Increase
(2) Maintenance Challenges
(3) Steeper Learning Curve

	Snowflake	Star
(1) ease of Maintenance	→ No redundancy and hence more easy to Maintain & storage.	→ has redudar data and hence less easy to maintain / change.
(2) ease of use	→ More Complex queries and hence less easy to understand	→ less Complex queries and easy to understand
(3) Query Performance	→ More no of table, More no of FKs, More execution time.	→ less no of table → less no of FKs → less execution time
(4) Joins	Higher no of joins	→ fewer joins
(5) when to use	when the dimension table is relatively big in size, it reduces space	→ when the dimension table Contain less no. of rows, use star schema

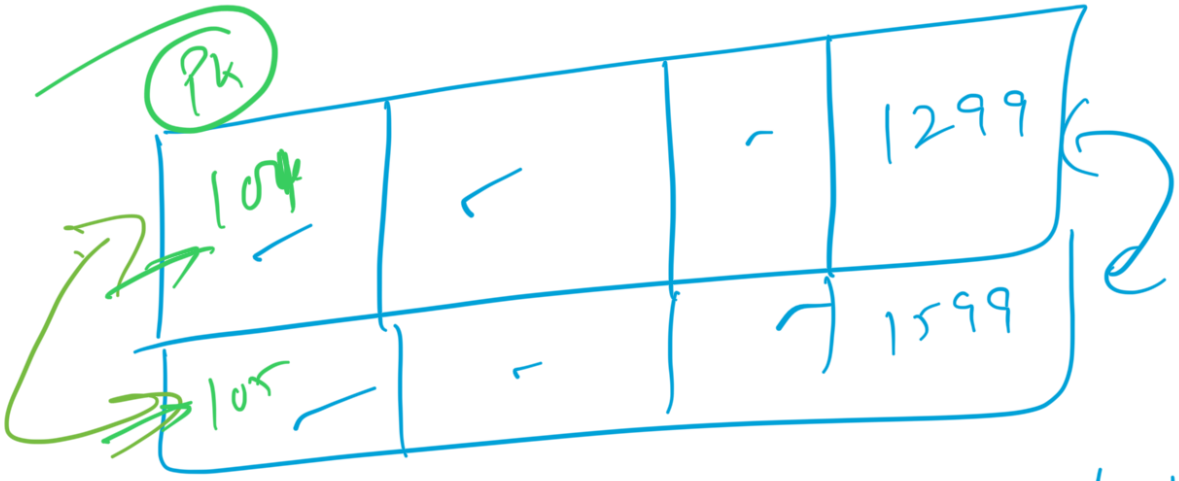
SCD → Slowly changing Dimension

→ used to analyze historical data



① SCD type 0 : Passive Method (No change is captured)

↳ In this method, no special action is performed upon dimensional changes

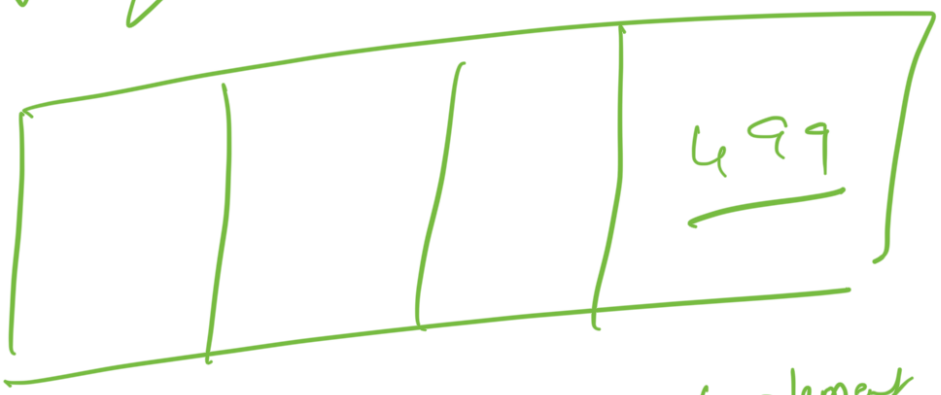


Pros :- no extra Consideration to designing DM.

Cons :- provide no control over recording & analyzing historical data

table changes.

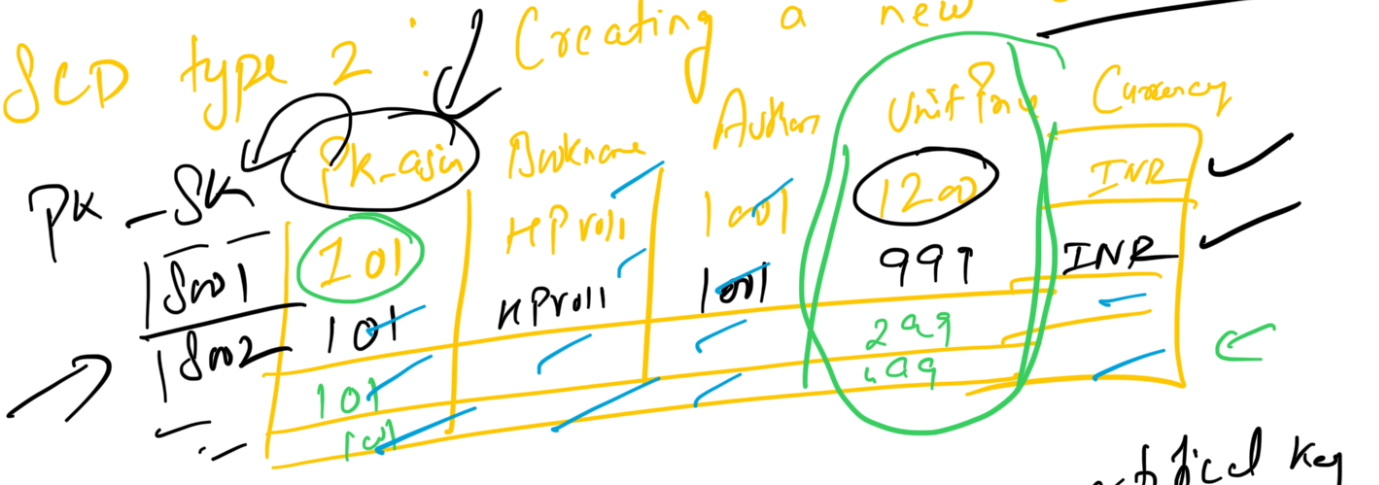
SCD Type 1 : overwriting the old values



Pros :- ↳ straightforward to implement
↳ we ignore old values
↳ uses less disk space since only 1 record exists.

Cons :- ↳ stores no historical data for analysis
↳ Hard to find previous state

SCD type 2 : Creating a new addition record.



Surrogate Keys → It is an artificial key

Surrogate key ✓ which is a unique identifier for a row used by different attributes of the same entity.

→ it is a system generated
→ it is not manipulable by a user.
→ it has no business context and is independent of domain.

Pros :- unlimited history
→ original table schema does not change.

Cons :- takes up more disk space.
↳ introduce change is very expensive.

SCD Type 3 :- Adding a new column

↳ An additional column available to capture the previous value of the attributes and the existing column presents the current value of the record by overwriting the old

value: Current Previous

Lag

Pros:- Suitable for tracking only the recent change.

↳ Disk Saved.

Cons:- It is by design limited to no of columns designated to store historical values.

Scd type 4:- Using historical table

Pros:- provides clear distinct b/w historical & Current data

↳ It resembles how an audit table store the changed data.

↳ Query on non-historical is fast.

Cons:- (1) Disk Space (slightly)

Cons :- ①

② Additional effort in ... Both tables.

Scd Type 6 :- (1, 2, 3)

Pros :- stores Complete picture of history in a 1 table.

→ Allows us to store multiple state changes while using type 3

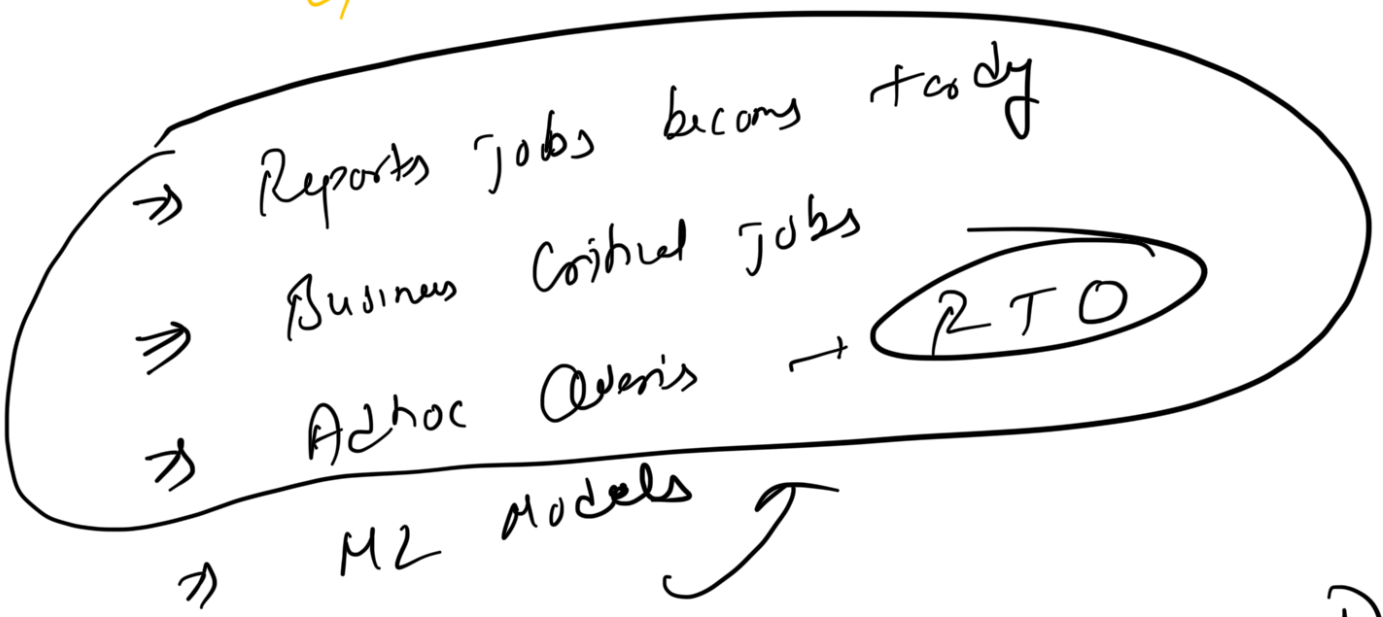
Cons :- more Complex to implement.

Airbnb

↳ 2014

→ Dashboarding / Reporting

↳ Adhoc Queries.



Solutions

↳ Read/write Replicas Dask

↳ cluster 2

Hive + Presto