Importing:-

Sqoop

→ Effective for Tables with Primary keys.
→ Uses Map/Reduce

Effective
→ Legacy Systems

Spark

→ Effective for Tables without Primary keys
→ Uses DAG
→ ~~Teradata~~

Table 2	
1	⅔ for 2018-11-08-05
2	⅔ for 2018-11-08-05
3	
4	

Added Values:

→ Using Spark can make Single Mapper Imports to Multiple Mappers [Reducing ingestion Time]
→ Go do Incroasment updates & custom
→ Ability Go do Incroasment updates [Flexibility] coding capability

Processing :- [Partitioned Parquet Tables]
Bucketed &
(or)
ORC

TimeStamps

2018-11-08-05 [For Hourly Refresh]

2018-11-08 [For Daily Refresh]

2018-11 [For Monthly Refresh]

partition2
partition1

Main_Table 1

1	2018-11-08-05
2	2018-11-08-05
3	2018-11-08-06
4	2018-11-08-05

Assuming Hourly Refresh

[Staging Script]

Catchup script to De Catchup inside xbit

→ Catchup script to fill a Date.

→ historical view of data [Descriptives Analytics]

→ time travel [Descriptives Analytics]

→ Schema Enforced & Downstream file/table format compatibility: [Effective Integration]

→ Partitioned [Reduces Reading Time]

Staging scripts clean up Added Values :-

ReProcess :- { Bucketted ORC(or) Parquet Tables}

⟹ Unique Record Filteration

↳ via joining the most partitions to the most recent previous table.

Unique_Table 2

1	2018 - 11 - 08 - 05 -
2	2018 - 11 - 08 - 06
4	2018 - 11 - 08 - 06

Added Values :-

⟹ Data De-duplication

⟶ Unified View

Cloud :-

Virtualized Clusters
with Admins
based on Projects
} Spin Up has needed

| Zeepline |
| Spark |
| Delta |

Delta → Just Adding a Jar file.

Added Values :-

Reduced Security Concerns
Effective Resource Utilization

Others :-

Run Zeepline on top of Spark using to browse the Delta Lake tables.