HandlingMissingDatain ETL

1. WhatarethemostcommonreasonsformissingdatainETLpipelines?
Ans.
• Sourcesystemissues→Datanotcapturedduetofaultysensors,humanerror,or incomplete forms.
• Schemachangesupstream→Columnnamesorstructureschange,breaking extraction.
• Integrationerrors→APIfailures,connectiontimeouts,ormismatchedformatsduring extraction.
• Datacorruption→Lossduringtransferbecauseofnetworklatencyorhardwarefailures.
• Workflowbottlenecks→Jobsfailmid-processduetoresourcelimitsorpoor orchestration.
• Businessrules→Certainfieldsintentionallyleftblank(e.g.,optionalattributes).

2. WhyisblindlydeletingrowswithmissingvaluesconsideredabadpracticeinETL?
Ans.
• Lossofvaluableinformation→Entirerecordsarediscardedevenifmostfieldsare valid.
• Biasinanalysis→Ifmissingnessissystematic(e.g.,certainregionsorcustomer groups), deletion skews results.
• Reducedsamplesize→Smallerdatasetsweakenstatisticalpowerandmodel accuracy.
• Businessimpact→Importantinsights(likewhydataismissing)arelost,whichcould itself be meaningful.

3. Q3.Explainthedifferencebetween:
• **Listwisedeletion**
• **Columndeletion**
Alsomentiononescenariowhereeachisappropriate. Ans.
ListwiseDeletion
• Removesentirerowsifanyrequiredfieldismissing.
• Preservesschemabutreducesdatasetsize.
• Scenario:Customersurveydatawheremultipleanswersaremissing→bettertodrop the whole record to avoid incomplete profiles.
ColumnDeletion
• Removesentirecolumnsiftheyhavetoomanymissingvalues.
• Preservesrowcountbutlosesthatvariable.

• Scenario:Adatasetwhere“MiddleName”ismissingfor95%ofcustomers→dropping the column avoids clutter without harming analysis.

4. Whyismedianimputationpreferredovermeanimputationforskeweddatasuchas income?

Ans.

• Meanimputation→Sensitivetoextremevalues(outliers).Inskeweddistributionslike income, a few very high salaries can inflate the mean, making it unrepresentative.

• Medianimputation→Morerobustbecausethemedianisthemiddlevalue,unaffected by outliers. It better reflects the “typical” case in skewed datasets.

• Example:Ifmostincomesarearound50,000butafewareabove1,000,000,themean will be distorted, while the median stays close to the majority.

5. Whatisforwardfillandinwhattypeofdatasetisitmostuseful?

Ans.

• ForwardFill(FFILL):
Atechniquewheremissingvaluesarereplacedwiththelastvalidobservationcarried forward.

• Usefulness:

o Bestsuitedfortime-seriesorsequentialdatasets(e.g.,monthlysales,stock prices, sensor readings).

o Assumescontinuity—thelastknownvalueisareasonableproxyuntilanew one appears.

• Example:
IfsalesdataforMarchismissingbutFebruaryhad12,000,forwardfillassignsMarch= 12,000 until April’s actual figure arrives.

6. WhyshouldflaggingmissingvaluesbedonebeforeimputationinanETL workflow?

Ans.

• Preservesinformationaboutmissingness→Thefactthatavalueismissingcanitself carry business meaning (e.g., customers not disclosing income).

• Avoidslossoftransparency→Onceimputationisapplied,youcan’tdistinguish between original and filled values unless flagged beforehand.

• Supportsbetteranalysis→Analystscanstudypatternsofmissingnessseparately(e.g., which regions or groups have more missing data).

• Improvesmodelaccuracy→Machinelearningmodelscanusetheflagasanadditional feature, helping them account for missingness bias.

7. Considerascenariowhereincomeismissingformanycustomers.Howcanthis missingness itself provide business insights?

Ans.

• Customerreluctance→Missingincomemayindicatecustomersareunwillingto disclose sensitive financial details, signaling trust or privacy concerns.

• Segmentbehavior→Certaindemographicsorregionsmayhavehighermissingness, revealing cultural or socio-economic differences.

• Producttargeting→Ifhigh-valuecustomersavoidsharingincome,itmaysuggestthey don’t see relevance in providing it, guiding product design or survey strategy.

• Operationalgaps→Consistentmissingnesscouldhighlightflawsindatacollection