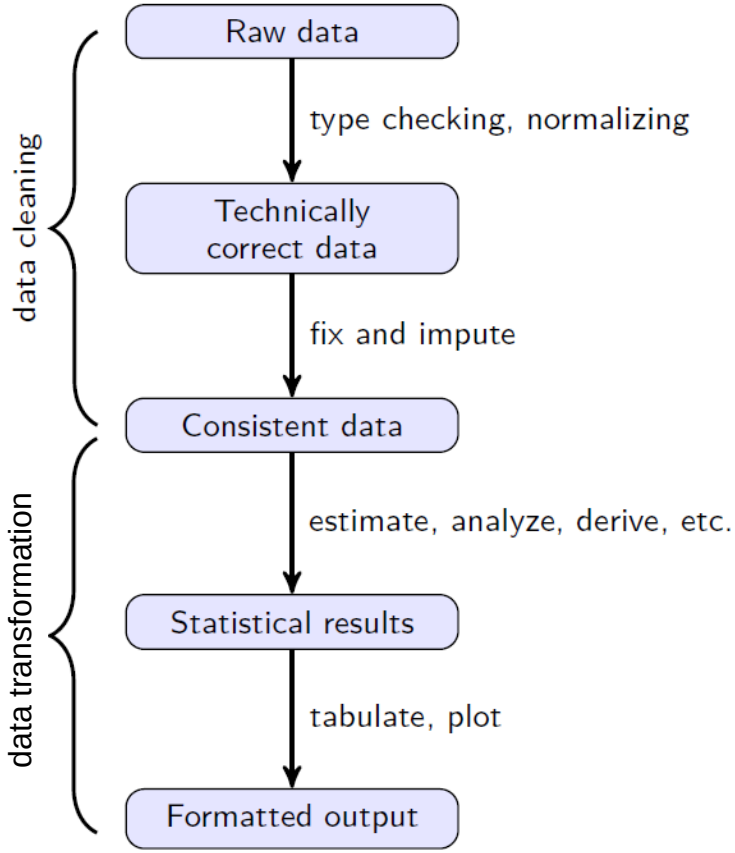
# Data Warehousing

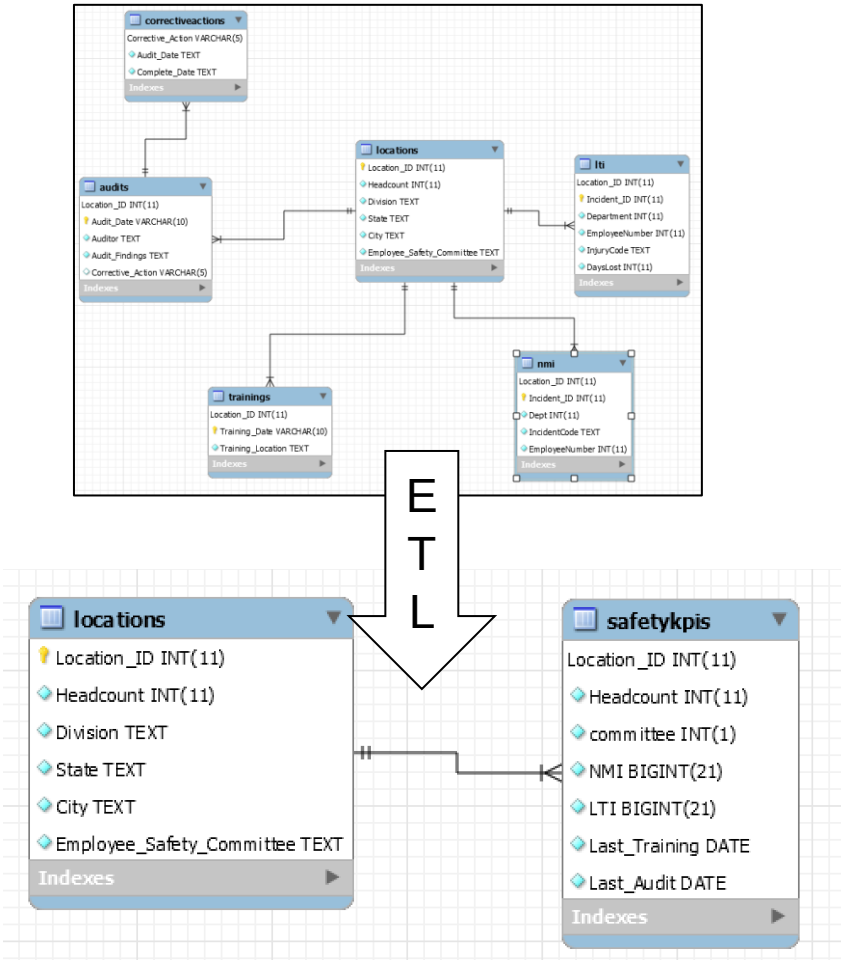

- Data Warehouse (DW) was proposed as a new type of database management system which would keep no transactional data but only summarized historical information for decision making purposes.
- W. H Inmon characterized a data warehouse as:
  - **“A subject-oriented, integrated, nonvolatile, time-variant collection of data in support of management’s decisions.”**

# DW Fact Table Creation

- Design of fact table
  - Granularity of fact table
  - Long vs. Wide
- Joining tables to create fact table – common issues
  - Many to many relationships
  - Outer joins
- Data Cleaning
  - Fix/delete bad records
  - Deduplication
  - Normalize units / Standardize fields for linking tables
  - Handle missing values, class imbalance, outliers
- Data Transformation/Aggregation
  - Summary fields
  - Calculated/Derived fields
  - Categories/Indicators



# Data Warehouse Example 1 – Safety KPIs




```
224  -- Data Warehouse - kpis view
225  CREATE OR REPLACE VIEW safetykpis AS
226  SELECT locations.Location_ID, Headcount,
227         IF(Employee_Safety_Committee = "yes", 1, 0) AS committee,
228         NMI, LTI, Last_Training, Last_Audit
229  FROM locations
230      JOIN    (   SELECT  Location_ID, COUNT(nmi.Incident_ID) AS NMI
231                  FROM nmi
232                  GROUP BY Location_ID
233              ) AS NMI_query
234          ON locations.Location_ID = NMI_query.Location_ID
235      JOIN    (   SELECT  Location_ID, COUNT(lti.Incident_ID) AS LTI
236                  FROM lti
237                  GROUP BY Location_ID
238              ) AS LTI_query
239          ON locations.Location_ID = LTI_query.Location_ID
240      JOIN    (   SELECT  Location_ID,
241                          MAX(STR_TO_DATE(Training_Date, "%m/%d/%Y")) AS Last_Training
242                  FROM trainings
243                  GROUP BY Location_ID
244              ) AS last_training
245          ON locations.Location_ID = last_training.Location_ID
246      JOIN    (   SELECT  Location_ID,
247                          MAX(STR_TO_DATE(Audit_Date, "%m/%d/%Y")) AS Last_Audit
248                  FROM audits
249                  GROUP BY Location_ID
250              ) AS last_audit
251          ON locations.Location_ID = last_audit.Location_ID;
```

See safetyDW.sql on Blackboard

# Fact Table Granularity

- Level of detail of one row of the fact able
- Higher (lower) granularity implies more (fewer) rows
- Trade-off between level of detailed analysis and storage requirements
- Examples:
  - One row of the fact table corresponds to one line on a purchase order
  - One row of the fact table corresponds to one purchase order
  - One row of the fact able corresponds to all purchase orders made by a customer

# Data Warehouse Example 2 – Online Retail

| InvoiceNo | Date | CustomerID | Country | Description | Quantity | UnitPrice | TotalAmount |
|---|---|---|---|---|---|---|---|
| 535789 | 2016-11-13 | 12428 | Finland | VINTAGE KITCHEN PRINT FRUITS | 3 | 5.06 | 15.18 |
| 535789 | 2016-11-13 | 12428 | Finland | RASPBERRY ANT COPPER FLOWER ... | 3 | 5.95 | 17.85 |
| 535789 | 2016-11-13 | 12428 | Finland | BLACK VINTAGE CRYSTAL EARRINGS | 4 | 3.75 | 15.00 |
| 535789 | 2016-11-13 | 12428 | Finland | DOORMAT KEEP CALM AND COME IN | 1 | 6.75 | 6.75 |
| 535789 | 2016-11-13 | 12428 | Finland | AMETHYST DIAMANTE EXPANDABLE ... | 3 | 4.25 | 12.75 |
| 535789 | 2016-11-13 | 12428 | Finland | PACK 20 DOLLY PEGS | 3 | 0.72 | 2.16 |
| 535789 | 2016-11-13 | 12428 | Finland | CANDY SPOT HAND BAG | 7 | 4.21 | 29.47 |
| 535789 | 2016-11-13 | 12428 | Finland | PICTURE DOMINOES | 2 | 1.45 | 2.9 |
| 535789 | 2016-11-13 | 12428 | Finland | I LOVE LONDON BEAKER | 4 | 1.25 | 5.0 |
| 535790 | 2017-08-22 | 18280 | United Kingdom | LILAC GAUZE BUTTERFLY LAMPSHADE | 1 | 0.42 | 0.4 |
| 535790 | 2017-08-22 | 18280 | United Kingdom | BLACK VINTAGE CRYSTAL EARRINGS | 4 | 3.75 | 15.00 |
| 535790 | 2017-08-22 | 18280 | United Kingdom | FOLK FELT HANGING MULTICOL GAR... | 1 | 7.62 | 7.62 |
| 535790 | 2017-08-22 | 18280 | United Kingdom | FIRST AID TIN | 7 | 3.25 | 22.75 |
| 535790 | 2017-08-22 | 18280 | United Kingdom | PINK DOG BOWL | 1 | 2.95 | 2.95 |
| 535790 | 2017-08-22 | 18280 | United Kingdom | LOLITA DESIGN COTTON TOTE BAG | 3 | 2.25 | 6.75 |
| 535791 | 2015-01-20 | 12673 | Germany | HOME SWEET HOME BOTTLE | 4 | 2.08 | 8.32 |
| 535791 | 2015-01-20 | 12673 | Germany | PINK ROSEBUD PEARL EARRINGS | 4 | 2.54 | 10.16 |
| 535791 | 2015-01-20 | 12673 | Germany | PACK/12 BLUE FOLKART CARDS | 7 | 2.95 | 20.65 |
| 535791 | 2015-01-20 | 12673 | Germany | EMBROIDERED RIBBON REEL SOPHIE | 6 | 3.75 | 22.50 |

ETL

| CustomerID | Country | CustomerTotal | CustomerAverage | DaysSinceLast | CustomerVolume | AvgGroup | LastGroup | VolumeGroup |
|---|---|---|---|---|---|---|---|---|
| 12431 | Australia | 2690.62 | 99.652593 | 168 | 27 | 1 | 3 | 1 |
| 12649 | Germany | 2221.35 | 100.970455 | 198 | 22 | 1 | 3 | 1 |
| 15480 | Malta | 1502.62 | 75.131000 | 208 | 20 | 5 | 4 | 1 |
| 12381 | Norway | 1514.64 | 75.732000 | 387 | 20 | 5 | 5 | 1 |
| 12673 | Germany | 1756.74 | 92.460000 | 9 | 19 | 2 | 1 | 1 |
| 12763 | Japan | 1646.78 | 86.672632 | 19 | 19 | 3 | 1 | 1 |
| 12758 | Portugal | 1498.59 | 78.873158 | 59 | 19 | 4 | 2 | 1 |
| [illegible] | Germany | 1423.08 | 79.060000 | 25 | 18 | 4 | 1 | 1 |
| [illegible] | Portugal | 1461.64 | 81.202222 | 48 | 18 | 4 | 1 | 1 |
| [illegible] | Germany | 1465.59 | 81.421667 | 110 | 18 | 4 | 2 | 1 |
| 12[illegible]6 | Portugal | 1455.83 | 80.879444 | 111 | 18 | 4 | 2 | 1 |
| 12444 | Norway | 1470.86 | 81.714444 | 446 | 18 | 4 | 5 | 1 |
| 12375 | Finland | 1448.25 | 85.191176 | 20 | 17 | 3 | 1 | 1 |
| 16321 | Australia | 1435.84 | 84.461176 | 68 | 17 | 3 | 2 | 1 |
| 12876 | Belgium | 1480.86 | 87.109412 | 74 | 17 | 3 | 2 | 1 |
| 12766 | Portugal | 1292.91 | 76.053529 | 74 | 17 | 5 | 2 | 1 |
| 12646 | USA | 1782.98 | 104.881176 | 145 | 17 | 1 | 3 | 1 |
| 12353 | Bahrain | 1210.98 | 71.234118 | 156 | 17 | 5 | 3 | 1 |
| 12352 | Norway | 1210.26 | 75.641250 | 8 | 16 | 5 | 1 | 1 |

See create_onlineretail.sql and 4_2_onlineRetail_RFM.sql on Blackboard