



# DATA SCIENCE AND BIG DATA ANALYTICS V2

**PARTICIPANT GUIDE**



## Dell Confidential and Proprietary

Copyright © 2018 Dell Inc. or its subsidiaries. All Rights Reserved. Dell, EMC and other trademarks are trademarks of Dell Inc. or its subsidiaries. Other trademarks may be trademarks of their respective owners.

# Table of Contents

|  |           |
|--|-----------|
| <b>Course introduction.....</b>                                  | <b>1</b>  |
| <b>Data Science and Big Data Analytics v2 .....</b>              | <b>2</b>  |
| Overall course goal .....  | 3         |
| Expected background .....  | 4         |
| Course objectives .....  | 5         |
| Prerequisite skills .....  | 6         |
| Course agenda .....  | 7         |
| <b>Introduction to Big Data analytics .....</b>                  | <b>9</b>  |
| <b>Lesson: Big Data and its characteristics .....</b>            | <b>11</b> |
| What are your thoughts on Big Data? .....                        | 12        |
| What are analysts' thoughts on Big Data? .....                   | 13        |
| What is Big Data? .....  | 14        |
| How significant is Big Data?.....                                | 16        |
| Characteristics of Big Data—(the 3 V's) .....                    | 18        |
| Characteristics of Big Data—volume.....                          | 19        |
| Characteristics of Big Data—velocity .....                       | 21        |
| Characteristics of Big Data—variety.....                         | 23        |
| Big Data characteristics—data structures .....                   | 26        |
| Big Data ecosystems .....  | 29        |
| Big Data ecosystem—data devices.....                             | 31        |
| Big Data ecosystem—data collectors.....                          | 33        |
| Big Data ecosystem—data aggregators.....                         | 34        |
| Big Data ecosystem—data users and buyers.....                    | 35        |
| Sources of Big Data .....  | 37        |
| Sources of Big Data—communication, media, and entertainment..... | 39        |
| Sources of Big Data—financial services.....                      | 40        |
| Sources of Big Data—healthcare .....                             | 41        |
| Sources of Big Data—social media .....                           | 42        |

|   |           |
|---|-----------|
| Sources of Big Data—Internet of Things (IoT).....   | 43        |
| Data repositories—an analyst perspective .....  | 44        |
| Data repositories—an analyst perspective—data island.....                                     | 46        |
| Data repositories—an analyst perspective—data warehouse.....                                  | 47        |
| Data repositories—an analyst perspective—analytic sandbox .....                               | 49        |
| Data repositories—an analyst perspective—data lake .....                                      | 50        |
| Concepts in practice—data lake with Dell EMC Isilon .....                                     | 52        |
| Why Big Data matters .....  | 54        |
| Mini-case study .....   | 57        |
| Check your knowledge.....   | 58        |
| Check your knowledge.....   | 59        |
| <b>Lesson: Business value from Big Data.....</b>  | <b>60</b> |
| Big Data analytics .....  | 61        |
| Business drivers to adopt Big Data analytics.....   | 63        |
| Deriving business value with Big Data analytics—communication, media, and entertainment ..... | 65        |
| Deriving business value with Big Data analytics—financial services .....                      | 67        |
| Deriving business value with Big Data analytics—healthcare.....                               | 69        |
| Data science—an emerging interdisciplinary field .....  | 70        |
| Business intelligence versus data science .....   | 71        |
| Business intelligence versus data science, cont.....  | 72        |
| Typical analytical architecture for business intelligence .....                               | 75        |
| Typical analytical architecture for business intelligence, cont .....                         | 78        |
| BI analytical architectures are not suitable for data science .....                           | 80        |
| Mini-case study—reducing customer churn at SuperMom&PopShop.....                              | 82        |
| Considerations for data science and Big Data analytics .....                                  | 83        |
| Check your knowledge.....   | 84        |
| <b>Lesson: Data scientist.....</b>  | <b>86</b> |
| Key roles for the new Big Data ecosystems .....   | 87        |
| Key roles for the new Big Data ecosystem—deep analytical talent .....                         | 89        |
| Key roles for the new Big Data ecosystem—data-savvy professionals.....                        | 91        |

|   |            |
|---|------------|
| Key roles for the new Big Data ecosystem—technology and data enablers ..... | 92         |
| Roles by technical and quantitative skills .....                            | 93         |
| Data scientist—an emerging career .....                                     | 94         |
| Profile of a data scientist .....   | 95         |
| Check your knowledge.....   | 97         |
| <b>Data analytics lifecycle.....</b>  | <b>99</b>  |
| <b>Lesson: Data analytics lifecycle overview.....</b>                       | <b>100</b> |
| Data analytics problems: What is your approach?.....                        | 101        |
| Key roles for a successful analytics project.....                           | 103        |
| Why use data analytic lifecycle .....                                       | 105        |
| Introduction to data analytic lifecycle.....                                | 106        |
| <b>Lesson: Discovery phase.....</b>   | <b>109</b> |
| Discovery phase—key activities.....   | 110        |
| Draft the business problem statement.....                                   | 111        |
| Discovery—interviewing the project sponsor .....                            | 112        |
| Discovery—interviewing other stakeholders and experts .....                 | 114        |
| Draft an analytics plan .....   | 115        |
| Analytics plan template .....   | 116        |
| Sales analytics project—using data analytics lifecycle .....                | 117        |
| Key discovery activities—sales analytics project.....                       | 118        |
| Key discovery activities—sales analytics project, cont. ....                | 119        |
| <b>Lesson: Data preparation phase .....</b>                                 | <b>120</b> |
| Data preparation—key activities.....  | 121        |
| Establish the analytic sandbox .....  | 122        |
| Extract, transform, load, and transform (ETLT) .....                        | 123        |
| Data exploration.....   | 124        |
| Data conditioning .....   | 125        |
| Summarize and visualize datasets .....                                      | 126        |
| Key data preparation activities—sales analytics project .....               | 127        |
| Key data preparation activities—attributes .....                            | 128        |

|   |            |
|---|------------|
| <b>Lesson: Model planning phase.....</b>                            | <b>130</b> |
| Model planning—key activities .....                                 | 131        |
| Variable selection.....   | 132        |
| Model selection.....  | 133        |
| Key model planning activities—sales analytics project.....          | 134        |
| <b>Lesson: Model Building Phase .....</b>                           | <b>135</b> |
| Model building—key activities .....                                 | 136        |
| Build training and test datasets .....                              | 137        |
| Train selected model.....   | 138        |
| Evaluate model.....   | 139        |
| Key model building activities—sales analytics project, 1 of 3 ..... | 140        |
| Key model building activities—sales analytics project, 2 of 3 ..... | 141        |
| Key model building activities—sales analytics project, 3 of 3 ..... | 142        |
| <b>Lesson: Communicate results phase.....</b>                       | <b>143</b> |
| Communicate results—key activities.....                             | 144        |
| Prepare presentation for sponsors and analysts .....                | 145        |
| Share project results with various audiences .....                  | 146        |
| Core deliverables to meet stakeholders needs.....                   | 148        |
| Key communicate results activities—sales analytics project .....    | 149        |
| <b>Lesson: Operationalize phase .....</b>                           | <b>150</b> |
| Operationalize—key activities .....                                 | 151        |
| Provide code and technical documentation.....                       | 152        |
| Onboard new team members.....                                       | 153        |
| Deploy model and monitor .....                                      | 154        |
| Key operations activities—sales analytics project .....             | 155        |
| <b>Lesson: Conclusion.....</b>                                      | <b>156</b> |
| Lifecycle continuation.....   | 157        |
| Concepts in practice—Pivotal Greenplum Database.....                | 158        |
| Check your knowledge.....   | 160        |

|   |            |
|---|------------|
| Check your knowledge.....                                       | 161        |
| Check your knowledge.....                                       | 162        |
| Module summary .....  | 163        |
| <b>Basic data analytics methods Using R .....</b>               | <b>164</b> |
| <b>Lesson: Introduction to the R programming language .....</b> | <b>165</b> |
| Introduction to R .....   | 166        |
| Basics of R programming.....                                    | 167        |
| Using the RStudio GUI.....                                      | 168        |
| Using the Help command in R.....                                | 169        |
| Importing data files into R .....                               | 170        |
| Importing database tables into R.....                           | 171        |
| Data types in R .....   | 172        |
| Attribute considerations in analytics .....                     | 174        |
| Common Data Structures in R .....                               | 176        |
| Vectors—atomic vectors and lists .....                          | 177        |
| Arrays .....  | 178        |
| Matrices .....  | 179        |
| Data frames .....   | 180        |
| Factors.....  | 181        |
| Exporting files and graphics out of R.....                      | 182        |
| Check your knowledge.....                                       | 183        |
| Lesson summary.....   | 184        |
| <b>Lesson: Analyzing and exploring data.....</b>                | <b>185</b> |
| What is data visualization?.....                                | 186        |
| Why is data visualization important? .....                      | 188        |
| Anscombe's Quartet .....  | 190        |
| Visualize before analyzing .....                                | 192        |
| Examining distribution of a single variable .....               | 193        |
| Density plots—what to look for.....                             | 194        |
| Evidence of dirty data .....                                    | 197        |
| Saturated data .....  | 198        |

|   |            |
|---|------------|
| Analyzing relationship between two variables .....                        | 200        |
| Two variables—what to look for .....                                      | 202        |
| Two variables—high-volume data: plotting .....                            | 203        |
| Establishing multiple pairwise relationships between variables .....      | 205        |
| Data exploration vs. presentation.....                                    | 207        |
| Check your knowledge.....   | 209        |
| Lesson summary.....   | 210        |
| <b>Lesson: Statistics for model building and evaluation .....</b>         | <b>211</b> |
| Statistical inference—drawing conclusions based on data .....             | 212        |
| Normal distribution.....  | 213        |
| Point estimation—normal distribution parameters, $\mu$ and $\sigma$ ..... | 214        |
| Confidence intervals .....  | 216        |
| Motivation for hypothesis testing .....                                   | 217        |
| The t-test on the mean ( $\mu=10$ ).....                                  | 218        |
| The t-test on the mean ( $\mu=9.7$ ).....                                 | 219        |
| Possible errors in hypothesis testing .....                               | 220        |
| Hypothesis tests for comparing multiple populations.....                  | 221        |
| Comparing Welch's t-test and Wilcoxon rank sum .....                      | 222        |
| Welch's t-test and Wilcoxon rank sum in R .....                           | 223        |
| Analysis of variance (ANOVA) .....  | 225        |
| Analysis of variance (ANOVA) in R .....                                   | 226        |
| Tukey honest significant differences .....                                | 227        |
| Statistics in data analytics lifecycle .....                              | 228        |
| Check your knowledge.....   | 229        |
| Lesson summary.....   | 230        |
| Module summary .....  | 231        |
| <b>Advanced analytics—theory and methods .....</b>                        | <b>232</b> |
| <b>Lesson: Introduction to advanced analytics—theory and methods.....</b> | <b>234</b> |
| Phase 3—model planning .....  | 235        |
| What kind of problem do I want to solve? How do I solve it? .....         | 236        |
| Why these analytic techniques?.....                                       | 239        |

|   |            |
|---|------------|
| <b>Lesson: K-means clustering .....</b>                         | <b>240</b> |
| K-means clustering .....  | 241        |
| Clustering .....  | 242        |
| K-means clustering—what is it? .....                            | 244        |
| K means clustering—use cases .....                              | 246        |
| Use-case example—online retailer .....                          | 248        |
| Algorithm .....   | 250        |
| Algorithm, cont.....  | 251        |
| Picking K.....  | 253        |
| Diagnostics—evaluating model.....                               | 255        |
| K-means clustering—reasons to choose (+) and cautions (-) ..... | 256        |
| Check your knowledge.....                                       | 258        |
| K-means clustering—summary .....                                | 260        |
| <b>Lesson: Association rules .....</b>                          | <b>261</b> |
| Association rules.....  | 262        |
| Grocery store—scenario .....                                    | 263        |
| Association rules.....  | 264        |
| Association rules—examples .....                                | 266        |
| Algorithm for association rules—Apriori .....                   | 267        |
| Apriori algorithm.....  | 268        |
| Apriori algorithm—support example .....                         | 269        |
| Apriori algorithm—confidence .....                              | 270        |
| Apriori algorithm—confidence example.....                       | 272        |
| Lift and leverage .....   | 273        |
| Apriori algorithm—lift and leverage example .....               | 274        |
| Sketch of algorithm .....                                       | 275        |
| Step 1—1-itemsets (L1) .....                                    | 276        |
| Step 2—2-itemsets (L2) .....                                    | 277        |
| Step 3—3-itemsets .....   | 278        |
| Finally—find confidence rules .....                             | 279        |
| Diagnostics .....   | 280        |
| Apriori—reasons to choose (+) and cautions (-) .....            | 281        |
| Check your knowledge.....                                       | 282        |

|   |            |
|---|------------|
| Association rules—summary.....                                | 283        |
| <b>Lesson: Linear regression .....</b>                        | <b>284</b> |
| Linear regression lesson topics.....                          | 285        |
| Regression.....   | 286        |
| Linear regression .....                                       | 287        |
| Linear regression model.....                                  | 288        |
| Example—linear regression with one input variable .....       | 289        |
| Representing categorical attributes .....                     | 291        |
| Fitting line with ordinary least squares (OLS) .....          | 293        |
| Interpreting estimated coefficients, $b_j$ .....              | 294        |
| Confidence and prediction intervals .....                     | 296        |
| Diagnostics—examining residuals .....                         | 297        |
| Diagnostics—plotting residuals .....                          | 298        |
| Diagnostics—residual normality assumption.....                | 299        |
| Diagnostics—using hold-out data.....                          | 300        |
| Diagnostics—other considerations .....                        | 301        |
| Linear regression—reasons to choose (+) and cautions (-)..... | 303        |
| Check your knowledge.....                                     | 304        |
| Linear regression—summary .....                               | 305        |
| <b>Lesson: Logistic regression .....</b>                      | <b>306</b> |
| Logistic regression topics.....                               | 307        |
| Logistic regression .....                                     | 308        |
| Logistic regression use cases .....                           | 309        |
| Logistic regression—technical description.....                | 310        |
| Logistic regression model—typical analysis steps.....         | 311        |
| Logistic regression—visualizing model.....                    | 312        |
| Diagnostics—confusion matrix .....                            | 313        |
| TPR and FPR are functions of threshold value .....            | 314        |
| Receiver operating characteristic (ROC) curve .....           | 315        |
| Diagnostics—plot histograms of scores .....                   | 316        |
| Diagnostic—sanity check coefficients .....                    | 317        |

|  |            |
|--|------------|
| Other diagnostics .....  | 318        |
| Logistic regression—reasons to choose (+) and cautions (-) ..... | 319        |
| Check your knowledge.....  | 320        |
| Logistic regression—summary .....                                | 321        |
| <b>Lesson: Text analysis.....</b>                                | <b>322</b> |
| Text analysis.....   | 323        |
| Text analysis, cont. ....  | 324        |
| Text analysis—problem-solving tasks .....                        | 325        |
| Example—term frequency.....                                      | 327        |
| Representing corpus—collection of documents and features.....    | 328        |
| Text classification—parsing and tokenizing.....                  | 330        |
| Extract and represent text .....                                 | 331        |
| Extract and represent text, cont. ....                           | 333        |
| Computing relevance—term frequency .....                         | 334        |
| Inverse document frequency (IDF) .....                           | 335        |
| More possibilities with text analysis.....                       | 336        |
| Natural language processing .....                                | 338        |
| Tough nut to crack—NLP .....                                     | 339        |
| Challenges—text analysis.....                                    | 340        |
| Check your knowledge.....  | 341        |
| Text analysis—summary .....                                      | 342        |
| <b>Lesson: Naïve Bayes .....</b>                                 | <b>343</b> |
| Naïve Bayes .....  | 344        |
| Classifiers .....  | 345        |
| Naïve Bayes classifier approach .....                            | 346        |
| Naïve Bayes—use cases .....                                      | 347        |
| Build training dataset to predict customer purchase .....        | 348        |
| Conditional probability.....                                     | 349        |
| Derivation of Bayes' Law.....                                    | 350        |
| Application of Bayes' Law .....                                  | 351        |
| Apply Naïve assumption and remove constant .....                 | 352        |
| Building Naïve Bayesian classifier .....                         | 353        |

|  |            |
|--|------------|
| Naïve Bayesian classifiers for product purchase example.....           | 354        |
| Naïve Bayesian classifier example, cont. ....                          | 355        |
| Naïve Bayesian implementation considerations .....                     | 356        |
| Diagnostics .....  | 358        |
| Naïve Bayesian classifier—reasons to choose (+) and cautions (-).....  | 359        |
| Check your knowledge.....  | 361        |
| Check your knowledge, cont. ....                                       | 363        |
| Naïve Bayesian classifiers—summary .....                               | 364        |
| <b>Lesson: Decision trees.....</b>                                     | <b>365</b> |
| Decision Trees.....  | 366        |
| Decision Tree classifier—what is it?.....                              | 367        |
| Decision Tree—example of visual structure .....                        | 369        |
| Decision Tree classifier—use cases.....                                | 370        |
| Example—credit risk problem .....                                      | 371        |
| General algorithm .....  | 373        |
| Step 1—Pick most informative attribute .....                           | 374        |
| Step 1—Pick most informative attribute—conditional entropy .....       | 375        |
| Step 1—which attribute is best classifier? .....                       | 376        |
| Step 1—which attribute is best classifier? (cont.) .....               | 377        |
| Conditional entropy example.....                                       | 378        |
| Steps 2 and 3—partition on selected variable .....                     | 379        |
| Diagnostics .....  | 380        |
| Decision Tree classifier– reasons to choose (+) and cautions (-) ..... | 381        |
| Which classifier should I try?.....                                    | 383        |
| Check your knowledge.....  | 384        |
| Decision trees—summary .....   | 386        |
| <b>Lesson: Time series analysis .....</b>                              | <b>387</b> |
| Time Series Analysis .....   | 388        |
| Time Series Analysis, cont.....  | 389        |
| Components of Time Series Analysis.....                                | 390        |
| Box-Jenkins method—what is it? .....                                   | 391        |

|   |            |
|---|------------|
| AR and MA models .....  | 392        |
| Time series—use cases.....  | 393        |
| Modeling time series .....  | 394        |
| Framework for ARIMA Time Series Modeling .....                              | 395        |
| Step 1—visualizing time series .....  | 396        |
| Step 2—stationarize series .....  | 397        |
| Step 2—stationarize series—detrending.....                                  | 398        |
| Step 2—stationarize series—seasonal adjustment .....                        | 399        |
| Step 3—plot ACF and PACF to identify optimal parameters .....               | 400        |
| Step 4—build model—ARMA (p, q).....   | 401        |
| Step 4—build model—ARIMA (p, d, q).....                                     | 402        |
| Step 4—build model—model selection.....                                     | 403        |
| Step 5—predict .....  | 404        |
| Time Series Analysis—reasons to choose (+) and cautions (-) .....           | 405        |
| Check your knowledge.....   | 406        |
| Time series analysis—summary .....  | 407        |
| Module summary .....  | 408        |
| <b>Advanced analytics—technology and tools .....</b>                        | <b>409</b> |
| <b>Lesson: Introduction to advanced analytics—technology and tools.....</b> | <b>411</b> |
| Challenges with Big Data beyond analytics.....                              | 412        |
| What is Apache Hadoop? .....  | 414        |
| Four main components of Apache Hadoop .....                                 | 416        |
| Hadoop Distributed File System.....   | 417        |
| Hadoop Distributed File System—assumption/goals.....                        | 419        |
| Hadoop Distributed File System—architecture.....                            | 421        |
| NameNode.....   | 422        |
| DataNodes.....  | 424        |
| Block replication.....  | 425        |
| Hadoop Distributed File System—file read.....                               | 426        |
| Hadoop Distributed File System—file write .....                             | 428        |
| Hadoop Distributed File System—not ideal in the following situations.....   | 430        |
| Introduction to MapReduce .....   | 431        |

|   |            |
|---|------------|
| What MapReduce is.....                                  | 432        |
| When to use MapReduce.....                              | 433        |
| MapReduce Steps .....                                   | 435        |
| MapReduce paradigm.....                                 | 436        |
| MapReduce—count words in document.....                  | 438        |
| Another word count example .....                        | 439        |
| Where MapReduce is used—some examples.....              | 440        |
| Example—social networking—eDiscovery .....              | 441        |
| Social triangle—first directed edge.....                | 443        |
| Social triangle—second directed edge .....              | 444        |
| Social triangle—third directed edge .....               | 445        |
| Hadoop operational modes .....                          | 446        |
| YARN—Yet Another Resource Negotiator .....              | 447        |
| YARN—architecture/components .....                      | 449        |
| Check your knowledge.....                               | 451        |
| Lesson—summary .....                                    | 452        |
| <b>Lesson: Hadoop ecosystem .....</b>                   | <b>453</b> |
| Lesson: Hadoop ecosystem .....                          | 454        |
| Key facets of Apache Hadoop ecosystem.....              | 455        |
| Apache projects covered in this lesson .....            | 456        |
| What is Pig?.....                                       | 457        |
| Writing Pig Latin.....                                  | 459        |
| Apache Hive .....                                       | 460        |
| Hive Shell and HiveQL.....                              | 462        |
| Temperature example—Hive .....                          | 463        |
| Hive comparison with SQL.....                           | 464        |
| HBase—the Hadoop database from Apache.....              | 466        |
| When to choose HBase .....                              | 468        |
| HBase comparison with traditional database.....         | 469        |
| Mahout.....   | 470        |
| Examples of useful algorithms available in Mahout ..... | 471        |
| Apache Spark™ .....                                     | 472        |

|  |            |
|--|------------|
| Spark uses memory instead of disk.....                                   | 474        |
| Sort competition.....  | 475        |
| Check your knowledge.....  | 476        |
| Lesson—summary .....   | 477        |
| <b>Lesson: In-database analytics SQL essentials .....</b>                | <b>478</b> |
| Lesson 3—in-database Analytics SQL essentials .....                      | 479        |
| Analytical databases and libraries.....                                  | 480        |
| Data computing appliance for Pivotal Greenplum—concepts in practice..... | 481        |
| Set operations.....  | 482        |
| Set operations—INTERSECT .....   | 483        |
| Set Operations—EXCEPT .....  | 484        |
| Set Operations—UNION ALL.....  | 485        |
| Set Operations—UNION .....   | 486        |
| Greenplum SQL OLAP grouping extensions .....                             | 487        |
| Standard GROUP BY example .....  | 488        |
| Standard GROUP BY example with UNION ALL .....                           | 489        |
| ROLLUP example .....   | 490        |
| GROUPING SETS example.....   | 491        |
| CUBE example .....   | 492        |
| GROUPING function example.....   | 493        |
| GROUP_ID function.....   | 494        |
| In-database text analysis .....  | 496        |
| Pattern matching—regular expressions (Regex) .....                       | 497        |
| Regular expression quantifiers.....                                      | 498        |
| Check your knowledge.....  | 499        |
| Lesson—summary .....   | 500        |
| <b>Lesson: Advanced SQL and MADlib .....</b>                             | <b>501</b> |
| Lesson—advanced SQL and MADlib .....                                     | 502        |
| Window functions.....  | 503        |
| Defining window specifications—OVER clause .....                         | 504        |
| RANK and ORDER BY .....  | 505        |
| Using OVER (ORDER BY...) clause .....                                    | 506        |

|  |            |
|--|------------|
| Window framing example.....                                    | 507        |
| Designating sliding window .....                               | 508        |
| Designating sliding window, cont. ....                         | 509        |
| Window framing example.....                                    | 510        |
| General syntax of window function.....                         | 511        |
| Built-in window functions.....                                 | 512        |
| Built-in window functions, cont. ....                          | 513        |
| Check your knowledge.....                                      | 514        |
| User-defined functions and aggregates.....                     | 516        |
| Anatomy of user-defined function.....                          | 517        |
| User-defined aggregates.....                                   | 519        |
| Ordered aggregates.....  | 521        |
| MADlib—definition .....  | 523        |
| MADlib in-database analytical functions.....                   | 524        |
| Calling MADlib functions—fast training, scoring .....          | 525        |
| MADlib—getting help .....                                      | 526        |
| Lesson—summary .....   | 527        |
| Check your knowledge.....                                      | 528        |
| Module summary—advanced analytics—technology and tools.....    | 529        |
| <b>Putting it all together .....</b>                           | <b>530</b> |
| <b>Lesson: Preparing to operationalize .....</b>               | <b>531</b> |
| Data analytics lifecycle.....                                  | 532        |
| Typical end-of-project scenario .....                          | 533        |
| Successful communication phase .....                           | 534        |
| Successful communication phase (cont. 1) .....                 | 535        |
| Successful communication phase (cont. 2) .....                 | 536        |
| Four core deliverables to meet stakeholders' needs .....       | 537        |
| Four core deliverables to meet stakeholders' needs, cont. .... | 538        |
| Considerations for technical documentation and code .....      | 539        |
| Common operationalize activities.....                          | 540        |
| Check your knowledge.....                                      | 542        |

|   |            |
|---|------------|
| Check your knowledge.....   | 543        |
| <b>Lesson: Preparing project presentations .....</b>                | <b>544</b> |
| Key aspects of project presentations.....                           | 545        |
| Key aspects of project presentations (cont. 1).....                 | 546        |
| Key aspects of project presentations (cont. 2).....                 | 547        |
| Key aspects of project presentations (cont. 3).....                 | 548        |
| Analytic plan for SuperMom&PopShop scenario.....                    | 549        |
| Analytic plan for SuperMom&PopShop scenario (cont.).....            | 551        |
| Developing core material for multiple audiences .....               | 553        |
| Components of final presentation—project goal.....                  | 554        |
| Components of final presentation—project goal (cont.).....          | 556        |
| Components of final presentation—executive summary.....             | 557        |
| Components of final presentation—executive summary (cont.).....     | 559        |
| Components of final presentation—approach .....                     | 560        |
| Components of final presentation—approach (cont.) .....             | 561        |
| Components of final presentation—model description .....            | 562        |
| Components of final presentation—key points with data .....         | 563        |
| Components of final presentation—model details .....                | 564        |
| Components of final presentation—model details (cont.) .....        | 565        |
| Components of final presentation—recommendation.....                | 566        |
| Summary of core components for target audiences .....               | 567        |
| Summary of core components for target audiences (cont.) .....       | 569        |
| Final considerations for preparing final presentation .....         | 571        |
| Final considerations for preparing final presentation (cont.) ..... | 572        |
| <b>Lesson: Data visualization techniques .....</b>                  | <b>573</b> |
| What is data visualization?.....                                    | 574        |
| Data visualization's importance.....                                | 576        |
| Common visualization tools .....                                    | 577        |
| Communicating key points supported with data .....                  | 578        |
| Key points supported with data .....                                | 579        |
| Key points supported with data—another example .....                | 580        |
| Iterative nature of building visualizations.....                    | 581        |

|   |     |
|---|-----|
| Iterative nature of building visualizations (cont.) ..... | 582 |
| Presenting pricing model results to sponsor.....          | 583 |
| Choosing correct chart type .....                         | 584 |
| Cleaning up graphic, example 1—before .....               | 586 |
| Cleaning up graphic, example 1—after .....                | 587 |
| Cleaning up graphic, example 2—before .....               | 589 |
| Cleaning up graphic, example 2—after .....                | 590 |
| Quick word about using 3D charts—avoid them.....          | 591 |
| Tips for building effective data visualizations.....      | 592 |
| Check your knowledge.....                                 | 593 |



## Course introduction



## Data Science and Big Data Analytics v2

## Overall course goal

### Overall course goal

- The goal of the Data Science and Big Data Analytics course is for you to be able to **immediately participate as a data science team member** on Big Data and other analytics projects.
  - Data scientist p-o-v
  - Open
  - Practical



DELL INC.

Here is the primary goal of the course. To achieve it, the course content is focused on the point of view (p-o-v) of a data scientist, it teaches concepts and principles in an open, vendor-neutral manner so they can be applied in any technology environment, and it provides many hands-on labs for practical experience with coaching from the instructor(s).

## Expected background

### Expected background

- Strong mathematical, quantitative capability
- Experience with statistical methods and basic proficiency with a statistical software package, such as R or RStudio, Minitab, Matlab, SAS, or SPSS
- Experience with the conditioning and management of business data including databases
- Basic programming skills, preferably including SQL



DELL INC.

The lectures in this course assume students have a strong numerate background with some experience of statistical software packages and the conditioning of business data. The labs in this course assume some programming expertise (preferably with R or SQL).

## Course objectives

### Course objectives

Upon successful completion of this course, participants should be able to:

- Immediately participate and contribute as a data science team member on Big Data and other analytics projects by:
  - Deploying the data analytics lifecycle to address Big Data analytics projects
  - Reframing a business challenge as an analytics challenge
  - Applying appropriate analytic techniques and tools to analyze Big Data, create statistical models, and identify insights that can lead to actionable results
  - Selecting appropriate data visualizations to clearly communicate analytic insights to business sponsors and analytic audiences
  - Using tools such as: R and RStudio, MapReduce/Hadoop, in-database analytics, Window and MADlib functions
- Explain how advanced analytics can be leveraged to create competitive advantage and how the data scientist role and skills differ from those of a traditional business intelligence analyst.



DELL INC.

Upon successful completion of this course, participants should be able to:

- Immediately participate and contribute as a data science team member on Big Data and other analytics projects by:
  - Deploying the data analytics lifecycle to address Big Data analytics projects
  - Reframing a business challenge as an analytics challenge
  - Applying appropriate analytic techniques and tools to analyze Big Data, create statistical models, and identify insights that can lead to actionable results
  - Selecting appropriate data visualizations to clearly communicate analytic insights to business sponsors and analytic audiences
  - Using tools such as: R and RStudio, MapReduce/Hadoop, in-database analytics, Window and MADlib functions
- Explain how advanced analytics can be leveraged to create competitive advantage and how the data scientist role and skills differ from those of a traditional business intelligence analyst.

## Prerequisite skills

### Prerequisite skills

- A strong quantitative background with a solid understanding of basic statistics, as would be found in a statistics 101 level course.
- Experience with a scripting language, such as Java, Perl, or Python (or R). Many of the lab examples taught in the course use R (with an RStudio GUI), which is an open source statistical tool and programming language.
- Experience with SQL (some course examples use PSQL).
- Experience with the conditioning and management of business data including databases.



To complete this course successfully and gain the maximum benefits from it, a student should have the following knowledge and skill sets:

- A strong quantitative background with a solid understanding of basic statistics, as would be found in a statistics 101 level course.
- Experience with a scripting language, such as Java, Perl, or Python (or R). Many of the lab examples taught in the course use R (with an RStudio GUI), which is an open source statistical tool and programming language.
- Experience with SQL (some course examples use PSQL).
- Experience with the conditioning and management of business data including databases.

## Course agenda

| Course agenda |  |  |
|---------------|--|--|
|               | Day 1  | Day 2  |
| AM            | Introduction to Big Data analytics                         | Basic data analytics methods using R                       |
|               |  | Basic data analytics methods using R lab                   |
|               | Lunch  | Lunch  |
| PM            | Data analytics lifecycle                                   | Advanced analytics – theory and methods (Session 1)        |
|               | Data analytics lifecycle lab                               | Advanced analytics – theory and methods (Session 1) labs   |
|               |  | Advanced analytics – theory and methods (Session 2)        |
|               |  | Advanced analytics – theory and methods (Session 2) labs   |
|               |  | Lunch  |
| AM            | Advanced analytics – theory and methods (Session 2)        | Advanced analytics – theory and methods (Session 2)        |
|               | Advanced analytics – theory and methods (Session 2) labs   | Advanced analytics – theory and methods (Session 2) labs   |
|               | Lunch  | Lunch  |
| PM            | Advanced analytics – technology and tools (Session 1)      | Putting it all together (Session 1)                        |
|               | Advanced analytics – technology and tools (Session 1) labs | Advanced analytics – technology and tools (Session 1) labs |
|               | Lunch  | Lunch  |
| AM            | Advanced analytics – technology and tools (Session 2)      | Advanced analytics – technology and tools (Session 2)      |
|               | Advanced analytics – technology and tools (Session 2) labs | Advanced analytics – technology and tools (Session 2) labs |
|               | Lunch  | Lunch  |
| PM            | Advanced analytics – technology and tools (Session 3)      | Putting it all together presentations                      |
|               | Advanced analytics – technology and tools (Session 3) labs | Advanced analytics – technology and tools (Session 3) labs |

| Course agenda, cont. |  |  |
|----------------------|--|--|
|                      | Day 4  | Day 5  |
| AM                   | Advanced analytics – theory and methods (Session 4)        | Advanced analytics – technology and tools (Session 3)      |
|                      | Advanced analytics – theory and methods (Session 4) labs   | Advanced analytics – technology and tools (Session 3) labs |
|                      | Lunch  | Lunch  |
| PM                   | Advanced analytics – technology and tools (Session 1)      | Putting it all together (Session 1)                        |
|                      | Advanced analytics – technology and tools (Session 1) labs | Advanced analytics – technology and tools (Session 1) labs |
|                      | Lunch  | Lunch  |
| AM                   | Advanced analytics – technology and tools (Session 2)      | Advanced analytics – technology and tools (Session 2)      |
|                      | Advanced analytics – technology and tools (Session 2) labs | Advanced analytics – technology and tools (Session 2) labs |
|                      | Lunch  | Lunch  |
| PM                   | Advanced analytics – technology and tools (Session 3)      | Putting it all together presentations                      |
|                      | Advanced analytics – technology and tools (Session 3) labs | Advanced analytics – technology and tools (Session 3) labs |

## Introductions

### Introductions



- Name
- Company
- Job Role
- Experience
- Expectations

1

INTRODUCTION

DELL INC.

# Introduction to Big Data analytics

## Introduction



### Introduction to Big Data analytics

Upon completing this module, you should be able to:

- ✓ Define Big Data and its characteristics.
- ✓ Identify the various sources of Big Data.
- ✓ Cite the business drivers for Big Data.
- ✓ Explain the evolving analytical architecture.
- ✓ Describe the role of data scientist.

This module covers the characteristics of Big Data, deriving business value from Big Data, and the emerging role of a data scientist.

Upon completing this module, you should be able to:

- Define Big Data and its characteristics.
- Identify the various sources of Big Data.
- Cite the business drivers for Big Data.
- Explain the evolving analytical architecture.
- Describe the role of data scientist.

## Lesson: Big Data and its characteristics

### Introduction

# Lesson: Big Data and its characteristics

DELL EMC

### Lesson: Big Data and its characteristics

In this lesson we discuss:

- The definition of Big Data
- Big Data characteristics and structure
- Sources of Big Data
- Understanding the business drivers for Big Data



DELL EMC

DELL EMC

This lesson covers the characteristics of Big Data and its sources.

## What are your thoughts on Big Data?

**Question / Discussion Topic:** What are your thoughts on Big Data?

What are your thoughts on Big Data?



Is there a threshold at which data becomes Big Data?  
How much does the complexity of its structure influence the designation as Big Data?  
Are you using any new or novel analytical techniques and tools to handle Big Data?

DATAFLYIC



### Discussion Notes:

## What are analysts' thoughts on Big Data?

What are analysts' thoughts on Big Data?

**Gartner**

**The Economist**

**Forbes**

- Information is the oil of the 21st century, and analytics is the combustion engine. – *Gartner*
- The world's most valuable resource is no longer oil, but data. – *The Economist*
- Big Data is a collection of data from traditional and digital sources inside and outside your company that represents a source for ongoing discovery and analysis. – *Forbes*

Consider these analysts' thoughts about Big Data:

- Information is the oil of the 21st century, and analytics is the combustion engine. – *Gartner*
- *The Economist* says that the world's most valuable resource is no longer oil as it was a century ago. It is data. These titans—Google, Amazon, Facebook—who are taking advantage of the data look unstoppable. They have enormous power.
- As per *Forbes*, Big Data is a collection of data from traditional and digital sources inside and outside your company that represents a source for ongoing discovery and analysis.

With all the interest surrounding Big Data and analytics, let us next examine the definition and characteristics of Big Data.

### References:

- [www.gartner.com/newsroom/id/1824919](http://www.gartner.com/newsroom/id/1824919)
- [www.cloudera.com/content/dam/www/static/documents/analyst-reports/idc-futurescape.pdf](http://www.cloudera.com/content/dam/www/static/documents/analyst-reports/idc-futurescape.pdf)
- [www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data/#732d65fb5c85](http://www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data/#732d65fb5c85)

## What is Big Data?

**What is Big Data?**

**Big Data:**  
Datasets so large they break traditional IT infrastructures



- Big Data not only signifies a huge volume of data, but also presents complex data types and structure, with an increasing volume of unstructured data.
- Data gets generated and changes rapidly, and also comes from diverse sources.

© 2014 Dell Inc. Dell EMC

There are many examples of emerging Big Data opportunities and solutions such as:

- Netflix suggesting your next movie rental.
- Dynamic monitoring of embedded sensors in bridges to detect real-time stresses and longer-term erosion.
- Hospitals predicting health ahead of time and reducing hospital admissions.
- Retailers analyzing digital video streams to optimize product and display layouts and promotional spaces on a store-by-store basis.

**Big Data:**  
Datasets so large they break traditional IT infrastructures



Big Data also exceeds the processing and analyzing capability of conventional IT infrastructure and software systems. It not only needs a highly scalable architecture for efficient processing and analysis, but also requires new and innovative technologies and methods for processing. These technologies typically use

## Lesson: Big Data and its characteristics

platforms such as distributed processing, massively parallel processing, and machine learning.

McKinsey Reference: [www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world](http://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world)

McKinsey & Co.; Big Data: The Next Frontier for Innovation, Competition and Productivity

## How significant is Big Data?

**How significant is Big Data?**

Every day:

|  |  |
|--|--|
|  Processes 0.5 petabytes  |  Generates 40 petabytes of transactional data |
|  Processes 24 petabytes   |  Touches 29 petabytes                         |
|  There are 413 petabytes produced by surveillance cameras around the world. |  |

1 petabyte = 1,000,000,000,000 bytes

© 2014 Dell Inc. All rights reserved. Dell EMC

As statistics show, data has grown from bytes, megabytes, gigabytes, and terabytes to petabytes (PB) in the last decade.

Every day:

- Facebook processes around 0.5 petabytes of data from feeds, updates, ads, and so on.
- Walmart generates 2.5 petabytes of transactional data.
- Google processes 24 petabytes of search data.
- The National Security Agency touches close to 29 PB of data for their day-to-day consumption and analysis.
- All the surveillance cameras around the world produce 413 PB of data.

Every day:



Processes 0.5 petabytes



Generates 40 petabytes of transactional data



Processes 24 petabytes



Touches 29 petabytes



There are 413 petabytes produced by surveillance cameras around the world.

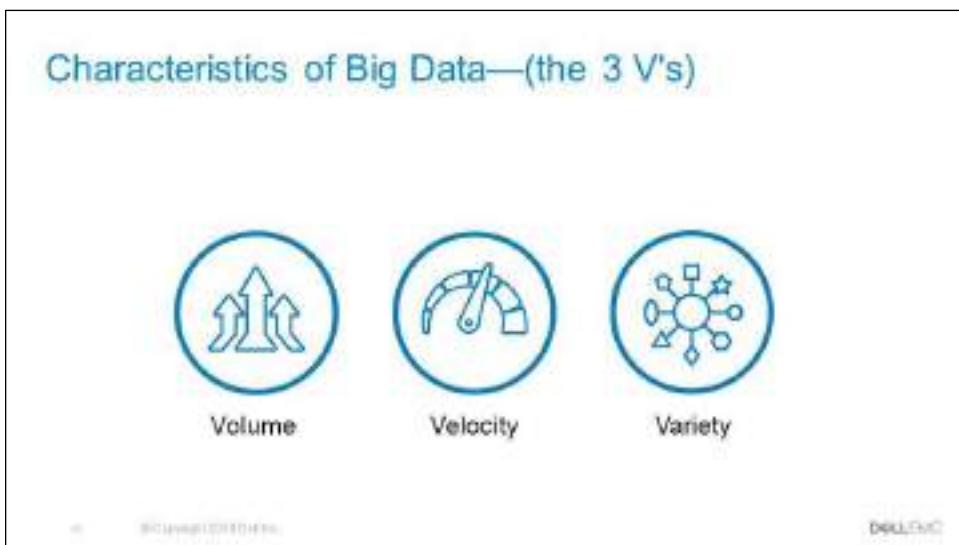


1 petabyte = 1,000,000,000,000,000 bytes

#### References:

- [www.cepro.com/article/video\\_surveillance\\_recordings\\_create\\_413\\_petabytes\\_of\\_data\\_every\\_day/](http://www.cepro.com/article/video_surveillance_recordings_create_413_petabytes_of_data_every_day/)
- [arstechnica.com/information-technology/2013/08/the-1-6-percent-of-the-internet-that-nsa-touches-is-bigger-than-it-seems/](http://arstechnica.com/information-technology/2013/08/the-1-6-percent-of-the-internet-that-nsa-touches-is-bigger-than-it-seems/)
- [www.slashgear.com/facebook-data-grows-by-over-500-tb-daily-23243691/](http://www.slashgear.com/facebook-data-grows-by-over-500-tb-daily-23243691/)
- [www.forbes.com/sites/bernardmarr/2017/01/23/really-big-data-at-walmart-real-time-insights-from-their-40-petabyte-data-cloud/](http://www.forbes.com/sites/bernardmarr/2017/01/23/really-big-data-at-walmart-real-time-insights-from-their-40-petabyte-data-cloud/)
- [datafloq.com/read/big-data-walmart-big-numbers-40-petabytes/1175](http://datafloq.com/read/big-data-walmart-big-numbers-40-petabytes/1175)

## Characteristics of Big Data—(the 3 V's)



## Characteristics of Big Data—volume

### Characteristics of Big Data—volume



- 2.5 quintillion bytes of data are created daily; 44x increase from 2009–2020.  
This would fill 10 million blue ray discs, the size of which would measure 4 Eiffel towers, one on top of another.
- An estimated 40 Zettabytes (43 trillion Gigabytes) of data will be created by 2020, an increase of 300 times from 2005. That is, 5,247 GB of machine data for every person on the planet.
- The population of the world is 7 billion; 6 billion people have cell phones; a source of huge volumes of data.

DELL INC.

## Overview

**Volume:** The word “big” in Big Data refers to the massive volumes of data. Organizations are witnessing an ever-increasing growth in data of all types, such as transaction-based data stored over the years, sensor data, and unstructured data streaming in from social media.

This growth in data is reaching Petabyte—and even Exabyte—scales. The excessive volume not only requires substantial cost-effective storage, but also creates challenges in data analysis. Self-driving cars will generate 2 PB of data. Some technological advancements such as wearables—including the Apple iwatch—generate enormous amounts of personal data that can help companies assess our health.

## Lesson: Big Data and its characteristics



It is estimated that 2.5 quintillion bytes of data are created around the world, and this number will double in 1.2 years. Apart from this data, there are Petabytes of data generated by commercial airlines, credit cards, transactions, and almost every sector, every second. Amazon sells 600 items per second. YouTube users view 12 years of

video in a single day, and the list goes on. This data would fill approximately 10 million blue ray discs, which, when stacked, would measure four Eiffel towers, one on top of another.

Estimates suggest that 40 Zettabytes of data will be created by 2020, an increase of approximately 300 times from 2005, most of which has been generated in the last couple of years.

The population of the world is 7 billion, and 6 billion people have cell phones that generate huge amounts of data that must be analyzed to understand and forecast customer needs.

## Characteristics of Big Data—velocity

Characteristics of Big Data—velocity



- Every 60 seconds, there are:
  - 98,000+ tweets.
  - 695,000 status updates on Facebook.
  - 698,445 Google searches.
- NYSE captures 1 TB of trade-related information during a trading session.
- The estimated rate of global Internet traffic by 2018 is 50,000 GB/sec.

DELL EMC

## Overview

**Velocity:** Velocity refers to the rate at which data is produced and changes, **and how the rapid generation of data must be processed to meet business requirements.**

Today, data is generated at an exceptional speed, and real-time or near-real-time analysis of the data is a challenge for many organizations. This data must be processed and analyzed, and the results delivered in a timely manner. An example of such a requirement is real-time *facial* recognition for screening passengers at airports.

## Lesson: Big Data and its characteristics

Consider these statistics that explain the rate at which data is being produced:

- Every 60 seconds, there are 98,000+ tweets; 695,000 status updates on Facebook; and hundreds of thousands of Google searches.
- The New York Stock Exchange captures 1 TB of trade-related information during a session. The biggest challenge is analyzing this data in real time, to generate portfolio and trade suggestions.



It is estimated that 50,000 GB/sec is the rate of global traffic by 2018. This traffic must be processed quickly to generate real-time insights.

This real-time data can help researchers and businesses make valuable decisions that provide strategic competitive advantages and ROI, if you are able to handle the velocity. Sampling this data efficiently can help deal with issues such as volume and velocity.

## Characteristics of Big Data—variety

**Characteristics of Big Data—variety**



- Data comes from social media in the form of tweets, feeds, status updates, and videos, structured and unstructured.
- Cisco estimates a total of 578 million wearables by 2019.
- As per estimate from VNI, wearables data traffic forecast for 2014 to 2019 will reach 292 EBs per year.
- Others varieties of data include data from:
  - Sensors in cars.
  - The healthcare industry.
  - Smart homes.
  - Air travel.

© 2014 Dell Inc. All rights reserved. Dell, the Dell logo, DELL.COM and DELL are trademarks of Dell Inc.

## Overview

**Variety:** Variety refers to the diversity in the formats and types of data. Data is generated by numerous sources in various structured and nonstructured forms.

Organizations face the challenge of managing, merging, and analyzing the different varieties of data in a cost-effective manner. The combination of data from various data sources and in various formats is a key challenge in Big Data analytics.

## Lesson: Big Data and its characteristics

Consider this example of combining many changing records of a particular patient with various published medical research to find the best treatment.

Varieties of data sources include:

- Clinical data from CPOE and clinical decision support systems—physician's written notes and prescriptions, medical imaging, laboratory, pharmacy, insurance, and other administrative data.
- Patient data in electronic patient records (EPRs).
- Machine-generated/sensor data, such as from monitoring vital signs.
- Social media posts, including Twitter feeds—so-called tweets—blogs, status updates on Facebook and other platforms, and web pages.
- Less patient-specific information, including emergency care data, news feeds, and articles in medical journals.



For the Big Data scientist, there is, among this vast amount and array of data, opportunity. By discovering associations and understanding patterns and trends within the data, Big Data analytics has the potential to improve care, save lives, and lower costs.

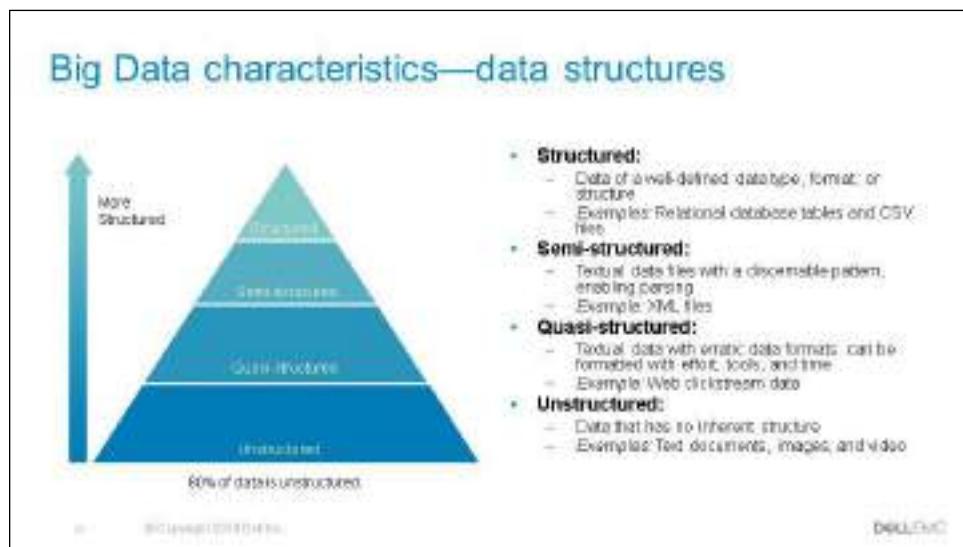
### References

- [investor.cisco.com/investor-relations/news-and-events/news/news-details/2015/Cisco-Visual-Networking-Index-VNI-Mobile-Forecast-Projects-Nearly-10-Fold-Global-Mobile-Data-Traffic-Growth-Over-Next-Five-Years/default.aspx](http://investor.cisco.com/investor-relations/news-and-events/news/news-details/2015/Cisco-Visual-Networking-Index-VNI-Mobile-Forecast-Projects-Nearly-10-Fold-Global-Mobile-Data-Traffic-Growth-Over-Next-Five-Years/default.aspx)

## Lesson: Big Data and its characteristics

- [www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html](http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html)
- [www.cbronline.com/internet-of-things/10-of-the-biggest-iot-data-generators-4586937/](http://www.cbronline.com/internet-of-things/10-of-the-biggest-iot-data-generators-4586937/)

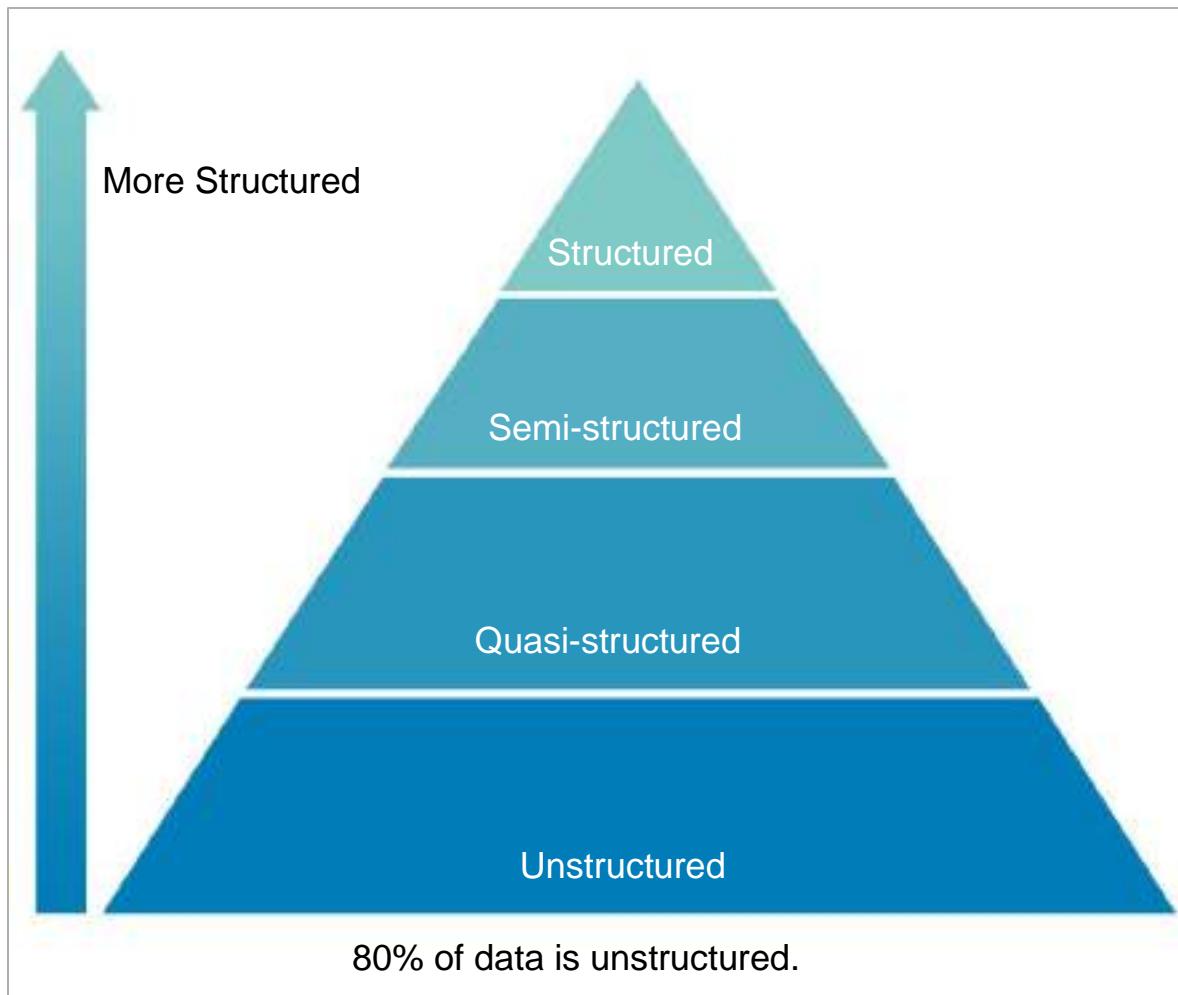
## Big Data characteristics—data structures



## Overview

The data structure of Big Data can be categorized into structured, semi-structured, quasi-structured, and unstructured. This graphic shows different types of data structures, with 80–90 percent of the future data growth coming from nonstructured data types—semi, quasi, and unstructured. The figure provides a description and examples of each type of data structure.

People tend to be most familiar with analyzing structured data, while semi-structured data, shown here as XML here, quasi-structured, shown here as a clickstream string, and unstructured data present different challenges and require different techniques to analyze.



Although the figure shows four different, separate types of data, in reality, these types can be mixed, at times. For instance, you may have a classic RDBMS storing call logs for a software support call center. In this case, you may have typical structured data such as date/time stamps, machine types, problem type, and operating system, which were probably entered by the support desk person from a pull-down menu GUI.

In addition, you may have unstructured or semi-structured data, such as free-form call log information, taken from an email ticket of the problem, or an actual phone call description of a technical problem and a solution. The most salient information is often hidden in there.

Another possibility would be voice logs or audio transcripts of the actual call that might be associated with the structured data. Until recently, most analysts would NOT be able to analyze the most common and highly structured data in this call log.

## Lesson: Big Data and its characteristics

history RDBMS, since the mining of the textual information is very labor-intensive and could not be easily automated.

## Big Data ecosystems

### Big Data ecosystems

As the new ecosystem takes shape, there are four main groups of players within this interconnected web:

- Data devices
- Data collectors
- Data aggregators
- Data users/buyers



© 2014 Dell Inc. All rights reserved.

DELL.COM

## Overview

Organizations and data collectors are realizing that the data they can gather from individuals contains intrinsic value and, as a result, a new economy is emerging. As this new digital economy continues to evolve, the market sees the introduction of data vendors and data cleaners that use crowdsourcing—such as Mechanical Turk and GalaxyZoo—to test the outcomes of machine learning techniques.



Other vendors offer added value by repackaging open-source tools in a simpler way and bringing the tools to market. Vendors such as Cloudera, Hortonworks, and Pivotal have provided this value-add for the open-source framework Hadoop.

## Lesson: Big Data and its characteristics

As the new ecosystem takes shape, there are four main groups of players within this interconnected web:

- Data devices
- Data collectors
- Data aggregators
- Data users/buyers

## Big Data ecosystem—data devices



### Data devices

Data devices—shown in section 1 of the figure—and the “Sensornet” gather data from multiple locations and continuously generate new data about this data. For each Gigabyte of new data created, an extra Petabyte of data is created about that data.



For example:

- **Online video games**

Consider someone playing an online video game through a computer, game

## Lesson: Big Data and its characteristics

console, or smartphone. In this case, the video game provider captures data about the skill and levels the player attained.

Intelligent systems monitor and log how and when the user plays the game. As a consequence, the game provider can fine-tune the difficulty of the game, suggest other related games that would most likely interest the user, and offer more equipment and enhancements for the character based on the user's age, gender, and interests.

This information may get stored locally or uploaded to the game provider's cloud to analyze the gaming habits and opportunities for upsell and cross-sell, and identify archetypical profiles of specific kinds of users.

- **Smartphones**

Smartphones provide another rich source of data. In addition to messaging and basic phone usage, they store and transmit data about Internet usage, SMS usage, and real-time location.

This metadata can be used for analyzing traffic patterns by scanning the density of smartphones in locations to track the speed of cars or the relative traffic congestion on busy roads. In this way, GPS devices in cars can give drivers real-time updates and offer alternative routes to avoid traffic delays.

- **Retail shopping loyalty cards**

Retail shopping loyalty cards record not just the amount an individual spends, but the locations of stores that person visits, the kinds of products purchased, the stores where goods are purchased most often, and the combinations of products purchased together. Collecting this data provides insights into shopping and travel habits and the likelihood of successful advertisement targeting for certain types of retail promotions.

## Big Data ecosystem—data collectors

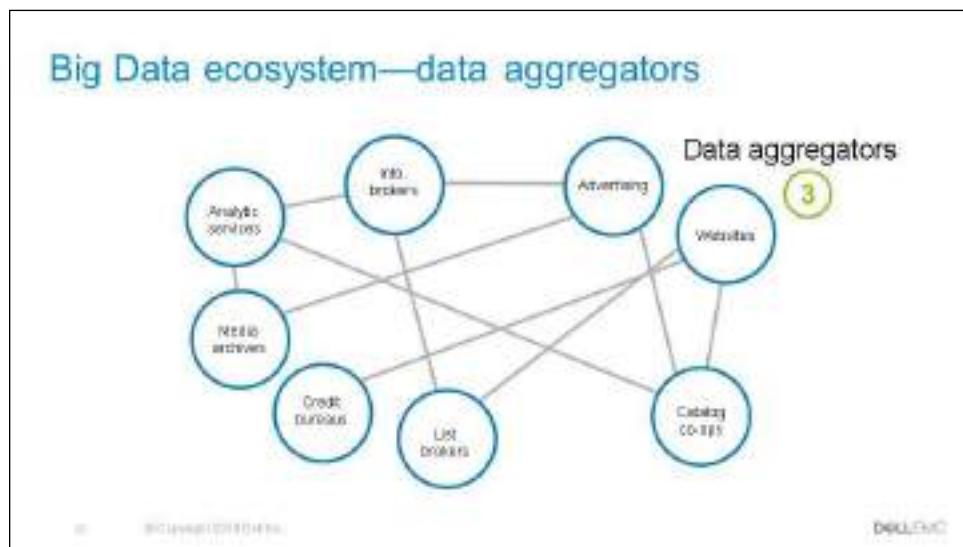


**Data collectors**—identified as 2 within the figure—include sample entities that collect data from the device and users.

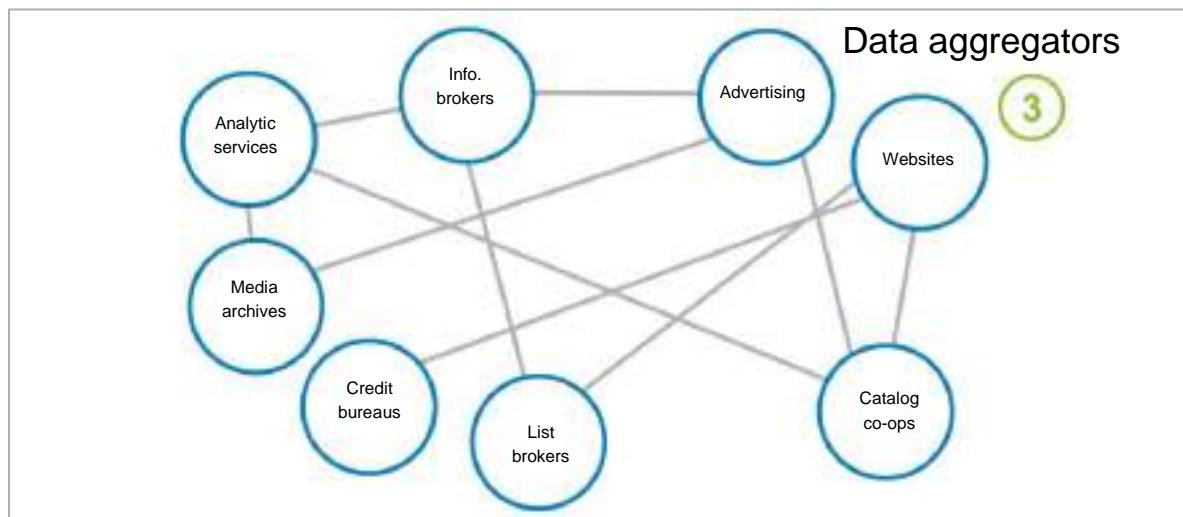


- Data results from a cable TV provider tracking the shows a person watches, which TV channels someone will and will not pay for to watch on demand, and the prices someone is willing to pay for premium TV content.
- Retail stores tracking the path a customer takes through their store while pushing a shopping cart with an RFID chip so they can gauge which products get the most foot traffic using geospatial data collected from the RFID chips.

## Big Data ecosystem—data aggregators

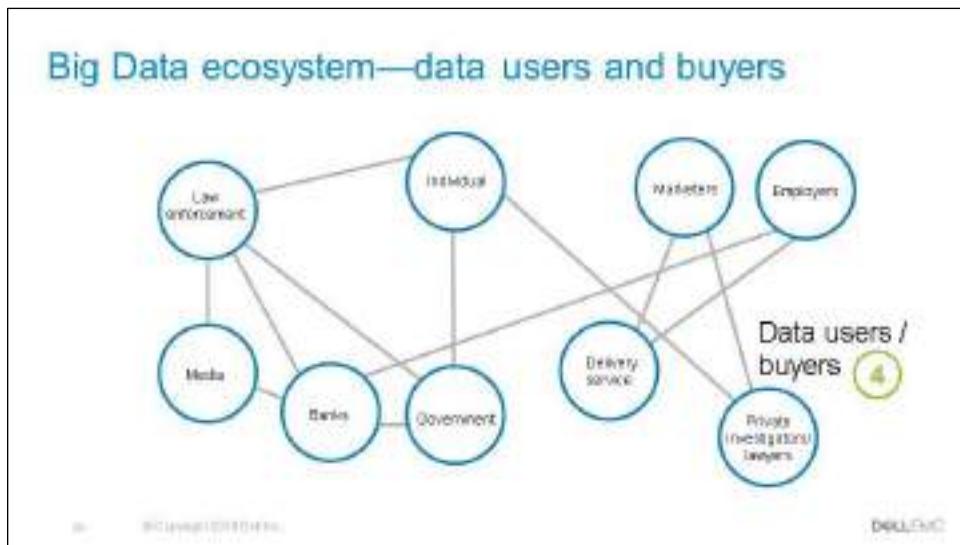


**Data aggregators**—marked as 3 in the figure—make sense of the data collected from the various entities from the “SensorNet” or the “Internet of Things.”

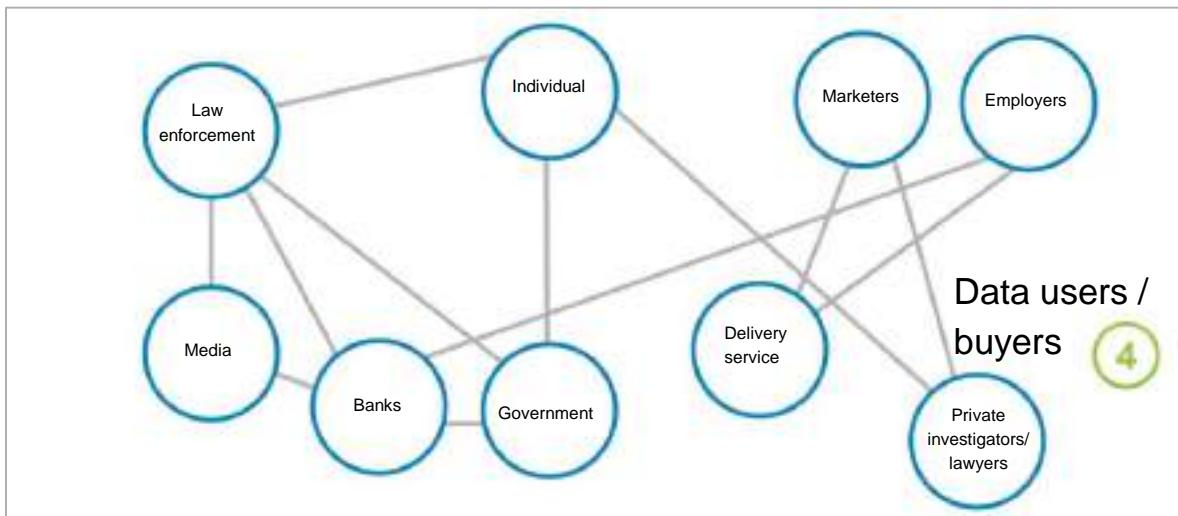


These organizations compile data from the devices and usage patterns collected by government agencies, retail stores, and websites. In turn, they can choose to transform and package the data as products to sell to list brokers. These brokers may want to generate marketing lists of people who may be good targets for specific ad campaigns.

## Big Data ecosystem—data users and buyers



**Data users and buyers** are denoted by (4) in the figure. These groups directly benefit from the data collected and aggregated by others within the data value chain.



- **Retail banks**

Retail banks, acting as a data buyer, may want to know which customers have the highest likelihood of applying for a second mortgage or a home equity line of credit. To provide input for this analysis, retail banks may purchase data from a data aggregator.

## Lesson: Big Data and its characteristics

This kind of data may include demographic information about people living in specific locations; people who seem to have a specific level of debt, yet still have solid credit scores—or other characteristics such as paying bills on time and having savings accounts—that can be used to infer credit worthiness; and those people who are searching the web for information about paying off debts or doing home remodeling projects.

Obtaining data from these various sources and aggregators enables a more targeted marketing campaign, which would have been more challenging before Big Data, due to the lack of information or high-performing technologies.

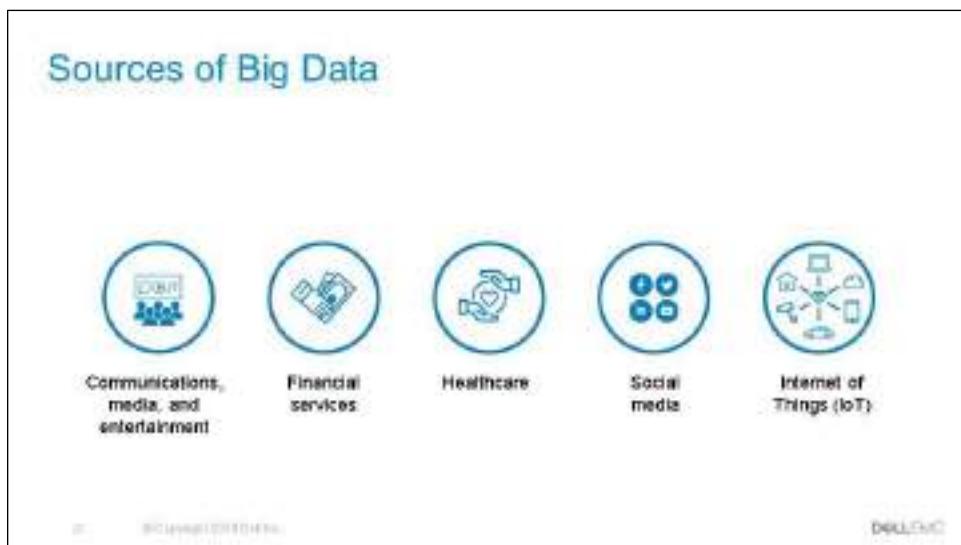
- **Technologies such as Hadoop**

Using technologies such as Hadoop to perform natural language processing on unstructured, textual data from social media websites, users can gauge the reaction to events such as presidential campaigns. People may, for example, want to determine public sentiments toward a candidate by analyzing related blogs and online comments.

Similarly, data users may want to track and prepare for natural disasters by identifying which areas a hurricane affects first and how it moves, based on which geographic areas are tweeting about it or discussing it via social media.

This emerging Big Data ecosystem illustrates that the kinds of data and the related market dynamics vary greatly. These datasets can include sensor data, text, structured datasets, and social media. With this variation in mind, it is worth recalling that these datasets do not work well within traditional enterprise data warehouses (EDWs), which were architected to streamline reporting and dashboards and be centrally managed. Instead, Big Data problems and projects require different approaches to succeed.

## Sources of Big Data



## Overview

Traditionally, data analytics solutions use data that is structured in rows and columns and use relational databases and data warehouses as sources of data. Big Data analytics has widened the scope of data sources to include both structured and nonstructured data and drive new value for organizations.



## Lesson: Big Data and its characteristics

Consider these Big Data sources, which are tapped across various industry verticals:

- Communication, media, and entertainment
- Financial services
- Healthcare
- Social media
- Internet of Things (IoT)

## Sources of Big Data—communication, media, and entertainment

Sources of Big Data—communication, media, and entertainment



- Customer feedback
- Contracts
- Network performance data
- Network traffic
- Network bandwidth usage
- User demographics
- Customer call records
- Social networks
- Viewing or usage habits

DELL EMC

For communication, media, and entertainment companies, the sources of Big Data may include information about customer engagement and customer satisfaction, contracts, network performance data, type of data and traffic passing over network, network bandwidth usage, user demographics, social network, viewing or usage habits, and customer call records.

## Sources of Big Data—financial services

**Sources of Big Data—financial services**



- Transaction records
- Trade messages
- World news
- Audio recordings
- Governance and regulatory data
- Customer feedback

Source: © Dell Inc. 2018

Big Data sources for financial services may include transaction records, trade messages, news, audio recording, governance and regulatory data, and customer feedback.

## Sources of Big Data—healthcare

Sources of Big Data—healthcare



- Genomic sequencing and diagnostic imaging
- Medical billing records
- Patient-specific data with socio-demographic
- Hospital care path
- Post-discharge information

Source: Dell EMC

Big Data sources for healthcare providers may be health records such as genomic sequencing and diagnostic imaging, medical billing records, patient-specific data with sociodemographic, hospital care path, and postdischarge information.

## Sources of Big Data—social media

Sources of Big Data—social media



- Facebook
- Twitter
- Emails
- Blogs
- LinkedIn
- WhatsApp
- YouTube

© 2018 Dell Inc.

For social media, the sources of Big Data may include Facebook, Twitter, LinkedIn, and YouTube, which contain information about various communities and individuals that allow the creation and sharing of ideas, career interests, and other forms of expression, which includes likes, dislikes, views, comments, and so on.

## Sources of Big Data—Internet of Things (IoT)

**Sources of Big Data—Internet of Things (IoT)**



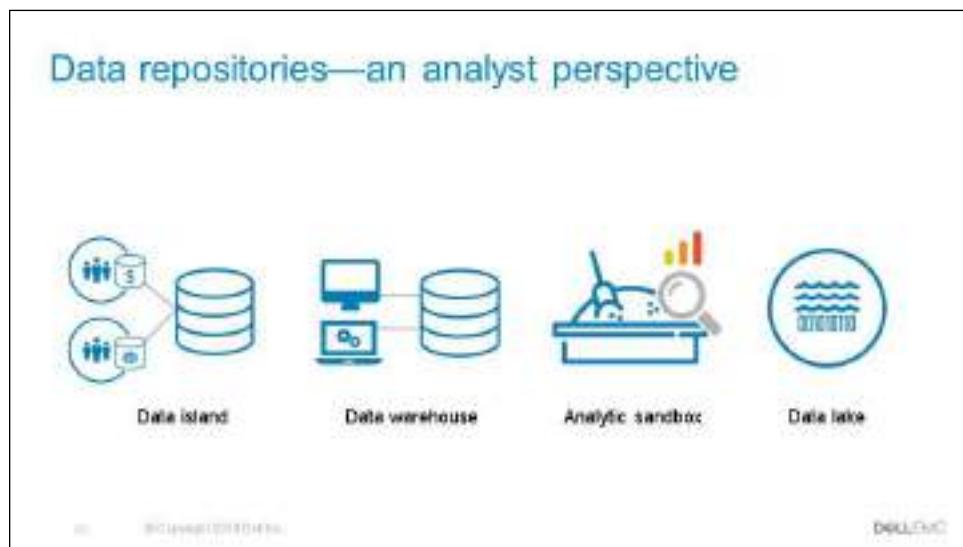
The diagram shows a central blue circle containing icons of a laptop, a smartphone, a car, a cloud, and a house, all connected by lines to various other icons representing different IoT devices like a television, a lightbulb, and a person. The background is dark green with a network of lines connecting the central hub to the periphery.

- Satellite communications
- Transmitters
- Receiver
- Tracking devices
- Smart phones
- Smart watches
- Public Web

© 2018 Dell Inc. All rights reserved. Dell, the Dell logo, and other Dell trademarks are trademarks of Dell Inc. All other trademarks are the property of their respective owners.

For the Internet of Things (IoT), the source of the Big Data may include satellite imaging, transmitter, receiver, tracking devices, and smart phones. The data collected from these global navigation satellite systems provides the geolocations and time information to the GPS receiver. Further, these systems provide critical positioning capabilities to military, civil, and commercial users.

## Data repositories—an analyst perspective



## Overview

Recent technology trends including the growth of third-platform applications and the Internet of Things (IoT) have generated an immense and growing wave of Big Data that requires the data storage and analytics platforms to scale significantly to handle the volume, velocity, and variety of this data. In addition, 80 percent or more of this data is unstructured and file-based data, which must be processed using advanced analytics methods and stored on modern storage infrastructure.

The data repositories Big Data uses include:

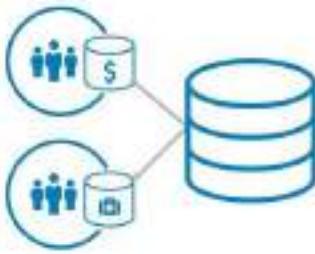
- Data island
- Data warehouse
- Analytic sandbox
- Data lake

## Lesson: Big Data and its characteristics



## Data repositories—an analyst perspective—data island

Data repositories—an analyst perspective—data island

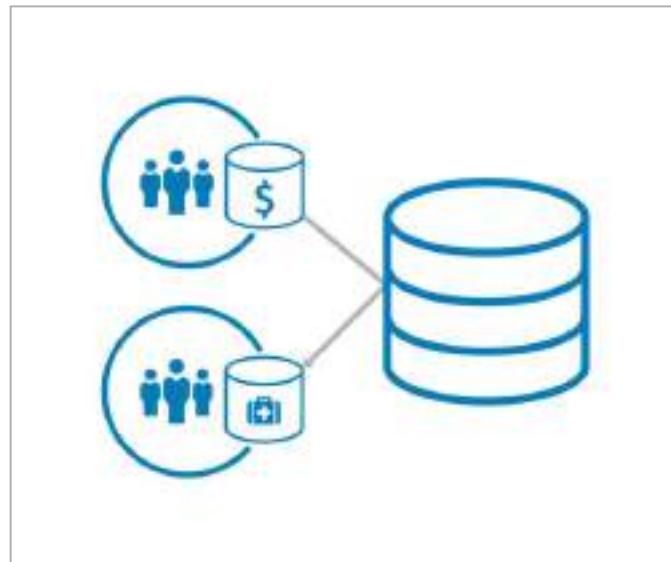


- Spreadsheets and low-volume DBs for recordkeeping
- Analyst dependent on data extracts

DATA ISLAND

The most ancient way of storing and analyzing data is to create a spreadsheet where data is structured in rows and columns and perform analysis to solve business problems. Users do not need heavy training as a database administrator to create spreadsheets, meaning business users could set these spreadsheets up quickly, independent of IT groups.

Proliferation of spreadsheets, however, may cause organizations to struggle with “many versions of the truth.” In other words, it was impossible to determine if you had the right version of a spreadsheet, with the most current data and logic in it. Moreover, if a user loses a laptop or it becomes corrupted, that is the end of the data and its logic. Many organizations still suffer from this challenge—Excel is still on millions of computers worldwide—which created the need for centralizing the data.



## Data repositories—an analyst perspective—data warehouse

Data repositories—an analyst perspective—data warehouse



- Critical for reporting and BI
- Data managed and controlled by IT groups and DBAs
- Often restrictions on analysts from building data sets

© 2018 Dell Inc. All rights reserved.

DELL.COM

### Overview

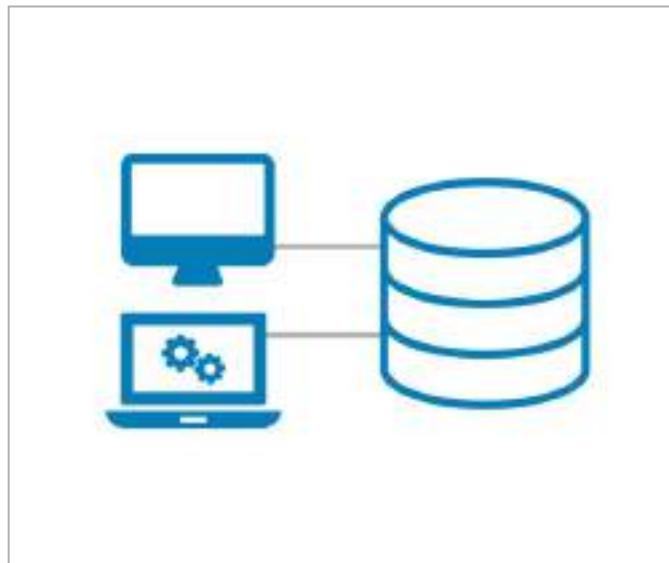
As data needs grew, companies such as Oracle, Teradata, and Microsoft—via SQL Server—offered more scalable data warehousing solutions. These technologies enabled the data to be managed centrally, providing benefits of security, failover, and a single repository where users could rely on getting an “official” source of data for financial reporting or other mission-critical tasks.

This structure also enabled the creation of online analytical processing (OLAP) cubes and business intelligence (BI) analytical tools, which provided users the ability to access dimensions within relational database management systems (RDBMS) quickly and find answers to streamline reporting needs.

Some providers also packaged more advanced logic and the ability to perform more in-depth analytical techniques such as regression and neural networks.

## Lesson: Big Data and its characteristics

Enterprise data warehouses, although critical for reporting and BI, tend to restrict the flexibility that a data analyst expects for performing robust analysis or data exploration. In this model, IT groups and database administrators (DBAs) manage and control data, and analysts must depend on IT for access and changes to the data schemas.



This tighter control and oversight also means longer lead times for analysts to get data, which generally must come from multiple sources. Another implication is that EDW rules restrict analysts from building datasets, which can cause shadow systems to emerge within organizations containing critical data for constructing analytic datasets, managed locally by power users.

## Data repositories—an analyst perspective—analytic sandbox

Data repositories—an analyst perspective—analytic sandbox



- Provides an area to merge and build datasets
- Enables rapid experimentation ("what if" analyses)
- Analyst-owned

Source: © 2018 Dell Inc.

DELL.COM

Analytic sandbox provides an area to merge and build datasets. This approach creates relationships to multiple data sources within an organization and saves the analyst the time of creating these data feeds on an individual basis. In addition, rather than the typical structured data in the EDW, analytic sandboxes can house a greater variety of data, such as raw data, textual data, and other kinds of unstructured data, without interfering with critical production databases.

## Data repositories—an analyst perspective—data lake

Data repositories—an analyst perspective—data lake



- Employs a "store-everything" approach
- Provides a foundation for Big Data analytics
- Ideally coupled with an analytic sandbox

DATA LAKES

A data lake is a collection of structured and nonstructured data assets that are stored as exact or near-exact copies of the source formats. The data lake architecture is a "store-everything" approach to Big Data.

Unlike conventional data warehouses, data is not classified when it is stored in the repository, as the value of the data may not be clear at the outset. The data is also not arranged as per a specific schema and is stored using an object-based storage architecture. As a result, data preparation is eliminated and a data lake is less structured compared to a data warehouse.

Data is classified, organized, or analyzed only when it is accessed. When a business need arises, the data lake is queried, and the resultant subset of data is then analyzed to provide a solution. The purpose of a data lake is to present an unrefined view of data to highly skilled analysts, and to enable them to implement their own data refinement and analysis techniques.



Ideally, a data lake is coupled with an analytics sandbox. Analytics Sandbox commonly pulls data from all layers of the data lake. It acts mainly as a playground for data scientists to conduct data experiments.

## Concepts in practice—data lake with Dell EMC Isilon



Dell EMC Isilon OneFS operating system provides the intelligence behind the Isilon scale-out Network Attached Storage (NAS) solutions. With powerful features and capabilities to optimize storage at the core of the data lake, OneFS combines the three layers of traditional storage architectures—file system, volume manager, and data protection—into one unified software layer, creating a single intelligent file system that spans all nodes within a cluster.



Isilon provides integrated support for industry-standard protocols including SMB, NFS, FTP, HTTP, OpenStack Swift, and HDFS, to provide an efficient, shared

## Lesson: Big Data and its characteristics

storage infrastructure for your data lake. This support allows users to consolidate data, cut costs, and accelerate results for a wide range of workloads.

Reference:

[www.dellemc.com/en-us/storage/isilon/onefs-operating-system.htm](http://www.dellemc.com/en-us/storage/isilon/onefs-operating-system.htm)

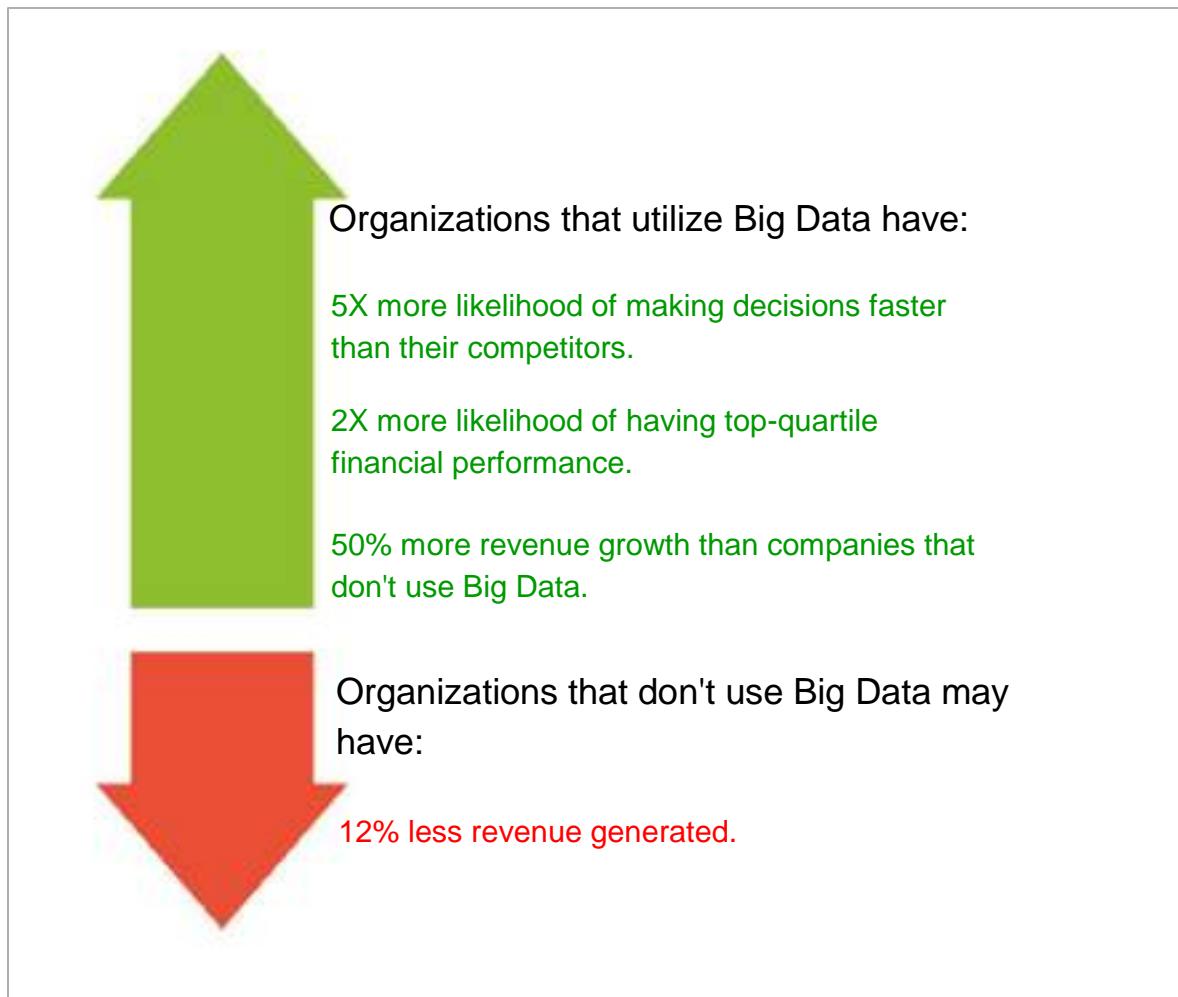
## Why Big Data matters



## Overview

Before proceeding further, it is important to understand why Big Data is useful and how it helps organizations. Let us find out.

Today, every organizational function from human resources to manufacturing **uses Big Data to drive decision-making.**



Estimates project that organizations that use Big Data have 50 percent more revenue growth than companies that do not use Big Data. They are two times more likely to have top-quartile performance and make decisions five times faster than their competitors.

It is also estimated that approximately 12 percent of revenues might be lost by financial services compared to their competitors, if they do not take advantage of Big Data.

**Big Data brings exciting new opportunities** that enable organizations to transform their businesses and rise above the competition. Organizations actively using Big Data have 50 percent higher revenue growth rates than those organizations that are not.

## Lesson: Big Data and its characteristics

There are many use cases of applying Big Data and analytics such as:

- Enhancing customer experience and sales by providing personalized recommendations
- Reducing Information storage costs by using Hadoop and cloud based analytics
- Detecting and preventing cybersecurity threats in real time
- Making decisions faster by analyzing real-time information

More details on the Global Technology Adoption Index 2015 can be found here:

[blog.dell.com/en-us/global-technology-adoption-index-2015/](http://blog.dell.com/en-us/global-technology-adoption-index-2015/).

References:

[www.bain.com/publications/articles/the-value-of-big-data.aspx](http://www.bain.com/publications/articles/the-value-of-big-data.aspx)

[www.gsam.com/content/gsam/global/en/market-insights/gsam-insights/gsam-perspectives/2016/big-data/infographic.html](http://www.gsam.com/content/gsam/global/en/market-insights/gsam-insights/gsam-perspectives/2016/big-data/infographic.html)

## Mini-case study

### Discussion Topic: Mini-case study

This lesson covered Big Data and its sources. Of course, collecting and storing Big Data is not the end goal; the focus must be on how to improve organizational outcomes. In the next lesson, you will see how analytics can be applied, but in the meantime, let us look at a mini-case study.

SuperMom&PopShop is a long-time brick and mortar retailer that also has a significant online retail business. However, as is the case with many traditional retailers, SuperMom&PopShop started its online presence separate from its physical store business and continues to run the two channels independently of each other.

Also, there are considerable competitive market pressures in the retail space, which continue to squeeze the profitability of the overall company.

### Mini-case study

**SuperMom&PopShop Retailer Scenario**

Traditional retailer with a somewhat independent online retail business

Significant market and operating cost pressures

From a business perspective, the challenges are:

- How to reduce customer churn?
- How to acquire new customers?
- How to best leverage the combined physical and online store businesses?

Mergers, acquisition, and store closures are expected industry-wide.

Dell EMC

## Check your knowledge

Check your knowledge

Which characteristic of big data refers to the diversity in the formats and types of data?

A. Variety      C. Value  
B. Variability    D. Volume

DATA SCIENCE

### Check your knowledge:

Which characteristic of Big Data refers to diversity in formats and types of data?

- A. Variety
- B. Variability
- C. Value
- D. Volume

### Question:

Which characteristic of Big Data refers to the diversity in the formats and types of data?

### Answer:

---

## Check your knowledge

Check your knowledge

Which data asset is an example of unstructured data?

A. News article text      C. Webserver log

B. XML data file      D. Database table

DELL EMC

### Check your knowledge:

Which data asset is an example of unstructured data?

- A. News article text
- B. XML datafile
- C. Webserver log
- D. Database table

### Question:

Which data asset is an example of unstructured data?

### Answer:

---

## Lesson: Business value from Big Data

### Introduction

# Lesson: Business value from Big Data

DELL EMC

### Lesson: Business value from Big Data

In this lesson we discuss:

- Business drivers for organizations to adopt Big Data analytics
- Business intelligence vs. data science
- Typical analytical architecture for business intelligence
- Considerations for Big Data analytics



CONTINUE

DELL EMC

## Big Data analytics

**Big Data analytics**



- Organizations can use their Big Data to:
  - Uncover new emerging trends.
  - Identify potential business opportunities.
  - Discover new ways to gain competitive advantages.
- Big Data demands an approach to analytics that is flexible, accessible, and fast.
- To maximize the value of Big Data, analysts:
  - Leverage data lakes that can store a massive amount of data.
  - Apply statistical and machine learning techniques.
  - Collaborate and share insights (it's a team sport).

© 2018 Dell Inc. All rights reserved. Dell EMC

## Overview

To harness the full power of Big Data, Big Data analytics are increasingly important. With Big Data analytics, organizations can use their Big Data to uncover new, emerging trends, identify potential business opportunities, and discover new ways to gain competitive advantages. In sum, Big Data analytics help organizations become more agile, identify opportunities, and respond faster.

## Lesson: Business value from Big Data



Big Data requires a modern platform that is optimized for high-performance analytics across both structured and nonstructured data. A query that takes 24 hours on a traditional data warehouse takes seconds on a modern analytics platform.

Moreover, a modern analytics platform has built-in advanced analytics and data mining services,

removing the lengthy process of copying data from a data warehouse into a specialized analytics database. As a result, a modern analytics platform delivers more accurate and timely insight to decision makers who have a direct impact to the business.

Big Data also demands an approach to analytics that is flexible, accessible, and fast. Traditional business intelligence tools provide sophisticated analytics and data mining capabilities; however, these tools tend to be rigid and hinder the analytical process.

With traditional tools, analysts must request from IT the desired datasets needed to answer a question, and IT must create a reporting environment in which the analysis can be performed. By the time the analysts are ready to perform the analysis, the supplied data is outdated.

Traditionally, the analysts work in isolation with fragmented datasets without tools that centralize and document insights to facilitate collaboration and knowledge sharing. As a result, results of an analysis are not fully optimized, and valuable insight is hidden and cannot be reused across the organization.

To maximize the value of Big Data, analysts must use data lakes that can store a massive amount of data from multiple sources, perform advanced analysis, collaborate and share insights, and iterate on the entire process continuously.

## Business drivers to adopt Big Data analytics

| Business drivers to adopt Big Data analytics     |   |
|--|---|
| Business driver                                  | Desired outcome   |
| Optimize business operations                     | Improve profitability and operating efficiency  |
| Identify business risk                           | Reduce customer churn and fraud   |
| Identify new business opportunities              | Increase sales revenue—for example, upsell, cross-sell, and find new customer prospects         |
| Stay informed of laws or regulatory requirements | Cost-effectively comply with industry regulations—anti-money laundering, Fair Lending, Basel II |

Here are four examples of common business problems that organizations contend with today, where they have an opportunity to apply advanced analytics to create competitive advantage. Rather than doing standard reporting on these areas, organizations can apply advanced analytical techniques to optimize processes and derive more value from these typical tasks.

| Business driver                                  | Desired outcomes  |
|--|---|
| Optimize business operations                     | Improve profitability and operating efficiency  |
| Identify business risk                           | Reduce customer churn and fraud   |
| Identify new business opportunities              | Increase sales revenue—for example, upsell, cross-sell, and find new customer prospects         |
| Stay informed of laws or regulatory requirements | Cost-effectively comply with industry regulations—anti-money laundering, Fair Lending, Basel II |

## Lesson: Business value from Big Data

The first three examples are not new problems—companies have been trying to reduce customer churn, increase sales, and cross-sell customers for many years. What is new is the opportunity to fuse advanced analytical techniques with Big Data to produce more impactful analyses for these old problems.

The fourth example listed here portrays emerging regulatory requirements. Many compliance and regulatory laws have been in existence for decades, but more requirements are added every year, which means more complexity and data requirements for organizations. These laws, such as anti-money laundering and fraud prevention, require advanced analytical techniques to manage well.

## Deriving business value with Big Data analytics— communication, media, and entertainment

### Deriving business value with Big Data analytics— communication, media, and entertainment



- Ability to predict what customer wants by analyzing usage patterns
- Ad Targeting to provide personalized advertising at the right time and right place
- Increased customer acquisition and retention by analyzing their social media behavior
- Efficient allocation of capital, to drive growth and profitability
- Enhanced planning and optimization of network services according to trends and predictive analytics

Let us revisit some of the earlier identified sources of Big Data and examine how analytics can be applied to derive business value.

Big Data analytics for communications, media, and entertainment: Service providers and media companies must harness massive amounts of data to improve user services. To be competitive, they must manage structured and unstructured data from multiple sources, consolidate information and make it actionable, and differentiate customer experience from that of the competition. Big Data analytics solutions can help them to transform data into actionable insight to improve customer experience and operational efficiency.



## Lesson: Business value from Big Data

When companies apply Big Data analytics, for example, they may:

- Predict customer usage patterns and provide accurate forecasting of the media content, network usage, and so on.
- Provide personalized advertising, so the dollars spent on advertising are used much more efficiently, and advertising is provided to customers at the right time and right place, increasing the probability of a sell.
- Increase customer retention and acquisition by analyzing customer sentiments and feedback on social media and taking proactive measures.
- Efficiently allocate capital to drive growth and profitability.
- Exercise enhanced planning and optimization of network services as per trends and predictive analytics.

## Deriving business value with Big Data analytics—financial services

Deriving business value with Big Data analytics—  
financial services



- Obtain a 360° view of customer to deliver better customer experience, improved branding, and increased revenues.
- Analyze call logs and social media activity to understand customer satisfaction levels and improve retention.
- Enhance Lender Risk management capability through behavioral analysis and understanding spending habits of customer.
- Use historical data to feed trading models and improve the performance of portfolio and revenue.
- Improve risk management by analyzing data from research, news, articles, social media, and so on.

DELL EMC

## Overview

Today's financial services organizations do not need to be convinced that Big Data is important. These organizations understand the concept of employing advanced analytics on real-time and historical data to gauge past performance and satisfy reporting and compliance requirements. They also employ data analytics to drive strategic business decisions, customize customer experience, and differentiate an organization from their competitors.

## Lesson: Business value from Big Data

Consider these examples related to the use of Big Data analytics in financial services:

- Financial institutions (FIs) with a global footprint can apply Big Data to develop a single view of a customer, which can promote delivery of an enhanced customer experience and in turn improve branding and increase revenues.
- FIs can analyze their internal call logs and social media activity to generate indications of customer dissatisfaction, allowing companies to act.
- Retail lenders and other FIs can use Big Data to analyze behavioral profiles of customers and understand their spending habits. In this manner, they can understand the customer better, to enhance risk management capability.
- FIs store large volumes of historical data from trading platforms. This data can be fed into trading models to improve performance of portfolio and revenue. Institutions can also use complex analytics and use magazines, references, and transaction data from different sources to generate insights.
- FIs improve risk management by analyzing data from research, news, articles, social media, and so on.



## Deriving business value with Big Data analytics—healthcare

Deriving business value with Big Data analytics—  
healthcare



- Reduced hospital admissions and re-admissions by identifying high-risk patients ahead of time through data analytics
- Sensors embedded into every technology, creating streams of data that, when analyzed, provide insights into patient health and behavior
- Efficient allocation of capital between R&D, clinical trials, research, and so on
- Effective use of genome sequencing to make personalized medical suggestions

DELL EMC

Let us revisit some of the earlier-identified sources of Big Data and examine how analytics can be applied to derive business value for healthcare providers.

Big Data analytics solutions meet rapidly evolving healthcare and clinical requirements by:

- Identifying high-risk patients ahead of time through analytics and reducing their admissions and readmissions.
- Taking advantage of sensors embedded in every technology that create streams of data to analyze patients' health and behavior.
- Efficiently allocating capital between R&D, clinical trials, research, and so on.
- Effectively using genome sequencing to make personalized medical suggestions.

## Data science—an emerging interdisciplinary field

### Data science—an emerging interdisciplinary field

- Data science combines several existing disciplines:
  - Statistics
  - Mathematics
  - Data visualization
  - Machine learning
  - Computer science
- This combination enables insights (data mining) as well as foresights (predictions).

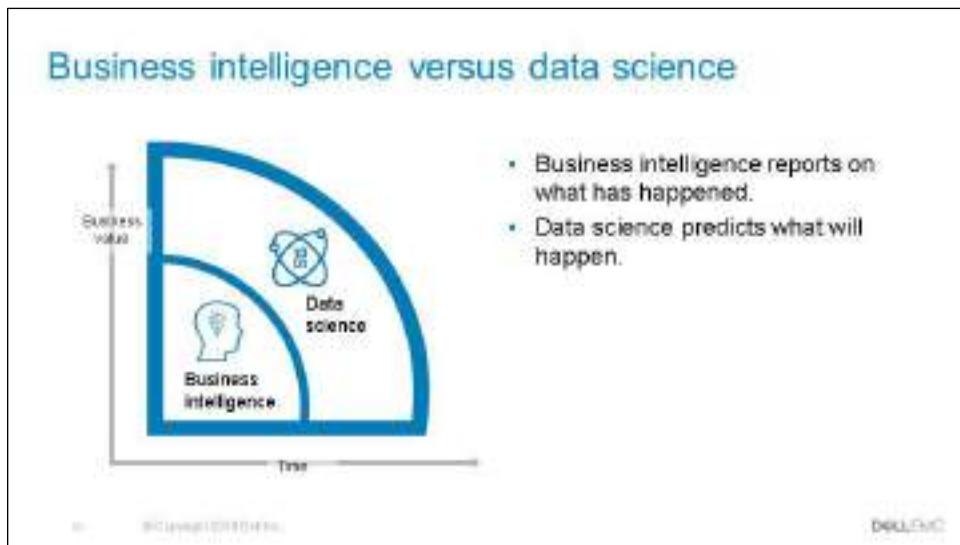
90

BIG DATA FOUNDATION

DELL EMC

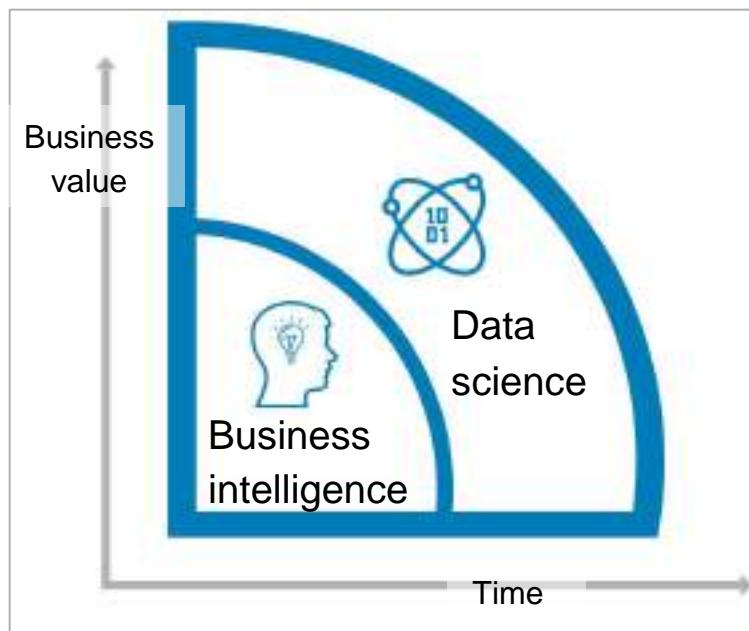
The emerging discipline of data science represents the combination of several existing disciplines, such as statistics, mathematics, data visualization, machine learning, and computer science, with the goal of extracting meaning from data. In simple terms, it is the umbrella of techniques used when trying to extract insights and foresights from data.

## Business intelligence versus data science



Let us compare data science to typical business reporting processes often known as “business intelligence.”

Business intelligence tends to provide reports, dashboards, and queries on business questions for the current period or in the past. BI systems make it easy to answer questions related to quarter-to-date revenue, progress toward quarterly targets, and understanding how much of a given product was sold in a prior quarter or year. These questions tend to be closed-ended and explain current or past behavior, typically by aggregating historical data and grouping it in some way. BI provides hindsight as well some insight, and generally answers questions related to “when” and “where” events occurred.



© Copyright 2018 Dell Inc.

## Business intelligence versus data science, cont.

**Business intelligence versus data science, cont.**

| Business intelligence  | Predictive analytics and data mining—data science  |
|--|--|
| A circular icon divided into four quadrants. The top-left quadrant contains a blue circle with a white 'D' and a small orange bar chart. The top-right quadrant contains a blue circle with a white 'B' and a small orange bar chart. The bottom-left quadrant contains a blue circle with a white 'D' and a small orange bar chart. The bottom-right quadrant contains a blue circle with a white 'B' and a small orange bar chart.   | A circular icon divided into four quadrants. The top-left quadrant contains a blue circle with a white 'D' and a small orange bar chart. The top-right quadrant contains a blue circle with a white 'B' and a small orange bar chart. The bottom-left quadrant contains a blue circle with a white 'D' and a small orange bar chart. The bottom-right quadrant contains a blue circle with a white 'B' and a small orange bar chart.   |
| <b>Business intelligence</b>   | <b>Predictive analytics and data mining—data science</b>   |
| <ul style="list-style-type: none"><li>+ Typical techniques and data types:<ul style="list-style-type: none"><li>- Standard and ad hoc reporting, dashboards, alerts, queries, data on demand</li><li>- Structured data, traditional sources, manageable datasets</li></ul></li><li>+ Common questions:<ul style="list-style-type: none"><li>- What happened last quarter?</li><li>- How many did we sell?</li><li>- Where is the problem? In which situations?</li></ul></li></ul> | <ul style="list-style-type: none"><li>+ Typical Techniques and Data Types:<ul style="list-style-type: none"><li>- Optimization, predictive modeling, forecasting, statistical analysis</li><li>- Structured/unstructured data; many types of sources, very large data sets</li></ul></li><li>+ Common Questions:<ul style="list-style-type: none"><li>- What if...?</li><li>- What's the optimal scenario for our business?</li><li>- What will happen next? What if these trends continue? Why is this happening?</li></ul></li></ul> |

DATAFLICO

By comparison, data science tends to use disaggregated data in a more forward-looking, exploratory way, focusing on assessing the present state and enabling informed decisions about the future.

Rather than aggregating historical data to look at how many of a given product sold in the previous quarter, a team may employ data science techniques to not only forecast future product sales and revenue more accurately, but to also be prescriptive in what raw materials should be ordered or what staffing levels should be.

|  |  |
|--|--|
|  |  |
| <b>Business intelligence</b>   | <b>Predictive analytics and data mining—data science</b>   |
| <ul style="list-style-type: none"> <li>• Typical Techniques and Data Types <ul style="list-style-type: none"> <li>– Standard and ad hoc reporting, dashboards, alerts, queries, details on demand</li> <li>– Structured data, traditional sources, manageable data sets</li> </ul> </li> <li>• Common Questions <ul style="list-style-type: none"> <li>– What happened last quarter?</li> <li>– How many did we sell?</li> <li>– Where is the problem? In which situations?</li> </ul> </li> </ul> | <ul style="list-style-type: none"> <li>• Typical techniques and data types <ul style="list-style-type: none"> <li>– Optimization, predictive modeling, forecasting, statistical analysis</li> <li>– Structured/unstructured data, many types of sources, very large datasets</li> </ul> </li> <li>• Common questions <ul style="list-style-type: none"> <li>– What if...?</li> <li>– What is the optimal scenario for our business?</li> <li>– What happens next? What if these trends continue? Why is this happening?</li> </ul> </li> </ul> |

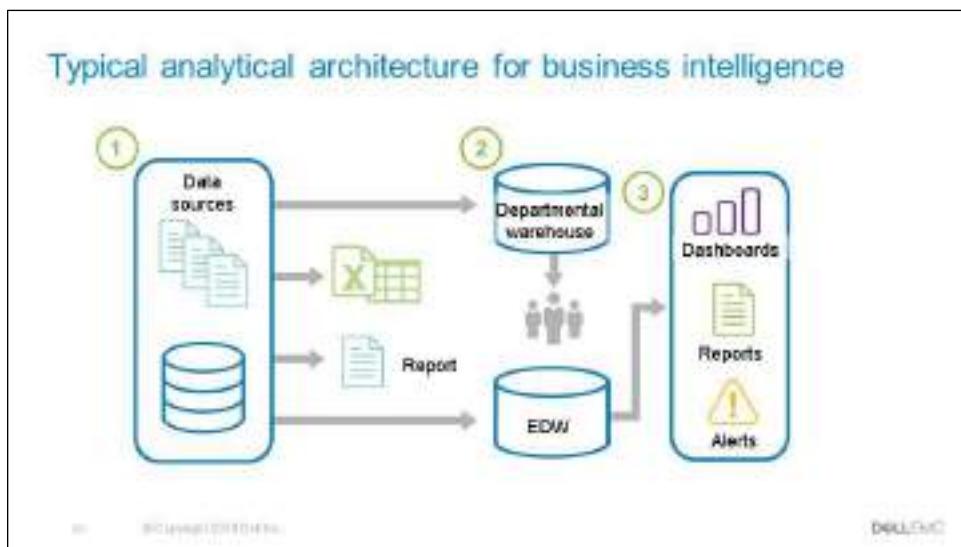
In addition, data science tends to be more exploratory in nature and may use scenario optimization to deal with more open-ended questions. This approach provides insight into current activity and foresight into future events, while generally focusing on questions related to “how” and “why” events occur.

Where BI problems tend to require highly structured data organized in rows and columns for accurate reporting, data science projects tend to use many types of data sources, including large or unconventional datasets.

## Lesson: Business value from Big Data

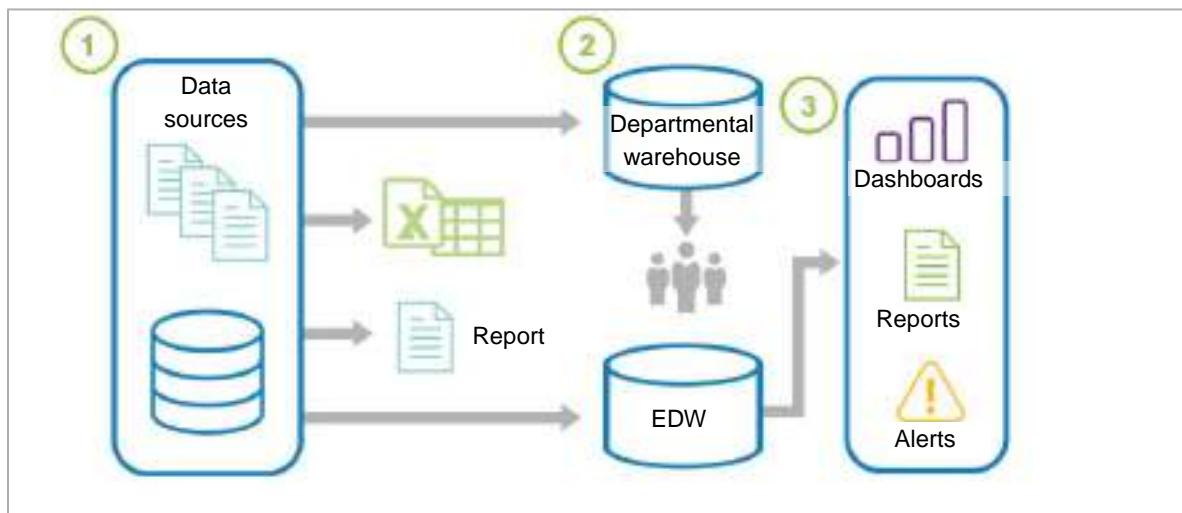
Depending on an organization's goals, its leaders may choose to embark on a BI project to do reporting, create dashboards, or perform simple visualizations. Or, they may choose data science projects if their objective is to do a more sophisticated analysis with disaggregated or varied datasets.

## Typical analytical architecture for business intelligence



## Overview

Data science projects need workspaces that are purpose-built for experimenting with data, with flexible and agile data architectures. Most organizations still have data warehouses that provide excellent support for traditional reporting and simple data analysis activities but unfortunately have a more difficult time supporting more robust analyses.



The graphic shows a typical data architecture and several of the challenges it presents to data scientists and others trying to do advanced analytics. This section

## Lesson: Business value from Big Data

examines the data flow to the data scientist and how this individual fits into the process of getting data to analyze on projects.

### Typical analytical architecture for business intelligence

#### 1. Step 1

For data sources to be loaded into the data warehouse, data must be well understood, structured, and normalized with the appropriate data type definitions. Although this kind of centralization enables security, backup, and failover of highly critical data, it also means that data typically must go through significant preprocessing and checkpoints before it can enter this sort of controlled environment. This preprocessing does not lend itself to data exploration and iterative analytics.

#### 2. Step 2

As a result of this level of control on the EDW, more local systems may emerge in the form of departmental warehouses and local data marts that business users create to accommodate their need for flexible analysis. These local data marts may not have the same constraints for security and structure as the main EDW and allow users to do some level of more in-depth analysis. However, these 1-off systems reside in isolation, often are not synchronized or integrated with other data stores, and may not be backed up.

#### 3. Step 3

After it is in the data warehouse, data is read by additional applications across the enterprise for BI and reporting purposes. These applications are high-priority operational processes getting critical data feeds from the data warehouses and repositories.

#### 4. Step 4

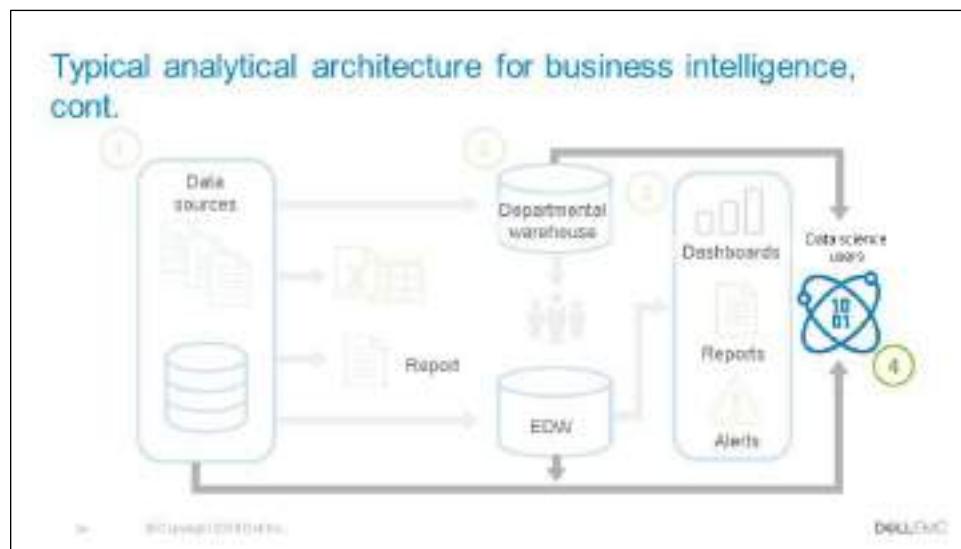
At the end of this workflow, data science users get data provisioned for their downstream analytics. Because users generally are not allowed to run custom or intensive analytics on production databases, analysts create data extracts from the EDW to analyze data offline in R or other local analytical tools.

Many times, these tools are limited to in-memory analytics on desktops analyzing samples of data, rather than the entire population of a dataset.

## Lesson: Business value from Big Data

Because these analyses are based on data extracts, they reside in a separate location, and the results of the analysis—and any insights on the quality of the data or anomalies—rarely are fed back into the main data repository.

## Typical analytical architecture for business intelligence, cont.



Because new data sources slowly accumulate in the EDW due to the rigorous validation and data structuring process, data is slow to move into the EDW, and the data schema is slow to change.



Departmental data warehouses may have been originally designed for a specific purpose and set of business needs, but over time evolved to house more data, some of which may be forced into existing schemas to enable BI and the creation of OLAP cubes for analysis and reporting.

Although the EDW achieves the objective of reporting and sometimes the creation

## Lesson: Business value from Big Data

of dashboards, EDWs generally limit the ability of analysts to iterate on the data in a separate nonproduction environment where they can conduct in-depth analytics or perform analysis on unstructured data.

## BI analytical architectures are not suitable for data science

BI analytical architectures are not suitable for data science



- High-value data is hard to reach and leverage.
- Data is moved from EDW to local analytical tools.
  - EDW may have masked/hidden meaningful data.
  - Sensitive data is stored on PCs.
  - This data is difficult to share and collaborate on.
- Isolated, ad hoc analytic projects, rather than centrally-managed harnessing of analytics.
- Preferred state: Analytic sandbox with access to the raw data.

DATA SCIENCE

The typical data architectures just described are designed for storing and processing mission-critical data, supporting enterprise applications, and enabling corporate reporting activities. Although reports and dashboards are still important for organizations, most traditional data architectures inhibit data exploration and more sophisticated analysis. Moreover, traditional data architectures have several additional implications for data scientists.

Because the EDWs are designed for central data management and reporting, those wanting data for analysis are generally prioritized after operational processes.

Data moves in batches from EDW to local analytical tools. This workflow means that data scientists are limited to performing in-memory analytics—such as with R, SAS, SPSS, or Excel—which restricts the size of the datasets they can use. As such, analysis may be subject to constraints of sampling, which can skew model accuracy.

Data science projects will remain isolated and ad hoc, rather than centrally managed. The implication of this isolation is that the organization can never harness the power of advanced analytics in a scalable way, and data science projects will exist as nonstandard initiatives, which are frequently not aligned with corporate business goals or strategy.

All these symptoms of the traditional data architecture result in a slow “time-to-insight” and lower business impact than could be achieved if the data were more readily accessible and supported by an environment that promoted advanced analytics. As stated earlier, one solution to this problem is to introduce analytic sandboxes to enable data scientists to perform advanced analytics in a controlled and sanctioned way. Meanwhile, the current data warehousing solutions continue offering reporting and BI services to support management and mission-critical operations.



## Mini-case study—reducing customer churn at SuperMom&PopShop

### Discussion Topic: Mini-case study

Returning to the SuperMom&PopShop scenario, one of the business challenges was reducing customer churn. It is often cheaper to retain a customer than acquire a new customer. How can the retailer better use Big Data analytics to reduce customer churn?

#### Mini-case study—reducing customer churn at SuperMom&PopShop

- Historical approach
  - Review quarterly reports generated on individual customer purchases.
  - In-home promotions are mailed to customers who have not made a purchase in a specified period of time.
- Big Data analytic approach
  - Build an analytical model to predict the likelihood of an individual customer churning.
  - Include new data sources and inputs.
  - Analyze previous types of purchases (home goods, tools, clothing, and so on).
  - Consider tendency to shop in-store, online, or both.
  - Consider distance from home to store locations.
  - Analyze customer demographics.

## Considerations for data science and Big Data analytics

Considerations for data science and Big Data analytics

- **Analysis flexibility**  
Where can the team explore and experiment with the data?
  - Data silos vs. analytic sandboxes
  - Analyst or IT owned
- **Decision making**  
How quickly must data-driven business decisions be made?
  - Batch processing
  - Real-time processing
- **Skills**  
Does the existing team have the necessary skills?
  - In-house expertise vs. outsourcing
  - Data science experts (aka data scientist)

Source: Dell EMC

Big Data projects carry with them several considerations that you must keep in mind to ensure this approach fits with what you are trying to achieve. Due to the characteristics of Big Data, these projects lend themselves to decision support for high-value, strategic decision-making with high processing complexity. The analytic techniques being used in this context must be iterative and flexible—analysis flexibility—due to the high volume of data and its complexity.

These conditions lead to complex analytical projects—such as predicting customer churn rates—that can be performed with some latency; consider the speed of decision-making needed. Or, these projects can be performed by operationalizing these analytical techniques using a combination of advanced analytical methods, Big Data, and machine learning algorithms to provide real-time—which requires high throughput—or near-real-time analysis, such as recommendation engines that look at your recent web history and purchasing behavior.

Also, the skill sets needed on the team are often different than what may have existed on business intelligence teams. Thus, the term data scientist has been coined to signify those individuals with deep analytical and strong technical skills.

## Check your knowledge

Check your knowledge

Which emerging discipline provides insights (data mining) and foresights (predictions) from data?

A. Topology      C. Complex analysis  
B. Business intelligence      D. Data science

DELL EMC

### Check your knowledge:

Which emerging discipline provides insights—data mining—and foresights—predictions—from data?

- A. Topology
- B. Business intelligence
- C. Complex analysis
- D. Data science

**Question:**

Which emerging discipline provides insights—data mining—and foresights—predictions—from data?

**Answer:**

---

## Lesson: Data scientist

### Introduction

Lesson: Data scientist

DELL EMC

### Lesson: Data scientist

This lesson covers:

- Key roles for the new Big Data ecosystem.
- Responsibilities of a data scientist.
- Profile of a data scientist.

Navigation icons: back, forward, search, etc.

DELL EMC

This lesson covers:

- Key roles for the new Big Data ecosystem.
- Responsibilities of a data scientist.
- Profile of a data scientist.

## Key roles for the new Big Data ecosystems



## Overview

As explained in the context of the Big Data ecosystem, new players have emerged to curate, store, produce, clean, and transact data. In addition, the need for applying more advanced analytical techniques to increasingly complex business problems has driven the emergence of new roles, new technology platforms, and new analytical methods. This section explores the new roles that address these needs.



## Lesson: Data scientist

The Big Data ecosystem demands three categories of roles described in the McKinsey Global study on Big Data, from May 2011:

- Deep analytic talent
- Data savvy professionals
- Technology and data enablers

## Key roles for the new Big Data ecosystem—deep analytical talent

Key roles for the new Big Data ecosystem—deep analytical talent



- Are tech-savvy with strong analytical skills
- Are able to build models and derive insights from data
- Are able to work with sandboxes
- Shortfall of some 250,000 data scientists
- Examples:
  - Statisticians
  - Data scientists

DATA SCIENCE

The first group—deep analytical talent—is technically savvy, with strong analytical skills. Members possess a combination of skills to handle raw, unstructured data and to apply complex analytical techniques at massive scales. This group has advanced training in quantitative disciplines, such as mathematics, statistics, and machine learning.

To do their jobs, members need access to a robust analytical sandbox or workspace where they can perform large-scale analytical data experiments. Examples of current professions fitting into this group include statisticians, economists, mathematicians, and the new role of the data scientist.

The recent McKinsey study estimates that the



## Lesson: Data scientist

number of graduates from data science programs could increase by a robust 7 percent per year, and our high-case scenario projects could have an even greater—12 percent—annual growth in demand, which would lead to a shortfall of some 250,000 data scientists. In addition, these estimates only reflect forecasted talent shortages in the United States; the number would be much larger on a global basis.

## Key roles for the new Big Data ecosystem—data-savvy professionals

Key roles for the new Big Data ecosystem—data-savvy professionals



- Are equipped with domain knowledge
- Have a basic knowledge of statistics or machine learning
- Are able to appreciate models built by data scientists
- Examples:
  - Financial analysts
  - Market research analysts
  - Business or functional manager

DELL EMC

The second group—data-savvy professionals—has less technical depth but has a basic knowledge of statistics or machine learning and can define key questions that can be answered using advanced analytics. These people tend to have a base knowledge of working with data, or an appreciation for some of the work data scientists and others with deep analytical talent perform. Examples of data-savvy professionals include financial analysts, market research analysts, life scientists, operations managers, and business and functional managers.

The McKinsey study forecasts the projected U.S. talent gap for this group to be 2 to 4 million people. Moving toward becoming a data-savvy professional is a critical step in broadening the perspective of managers, directors, and leaders, as this step provides an idea of the kinds of questions that can be solved with data.

## Key roles for the new Big Data ecosystem—technology and data enablers

Key roles for the new Big Data ecosystem—technology and data enablers



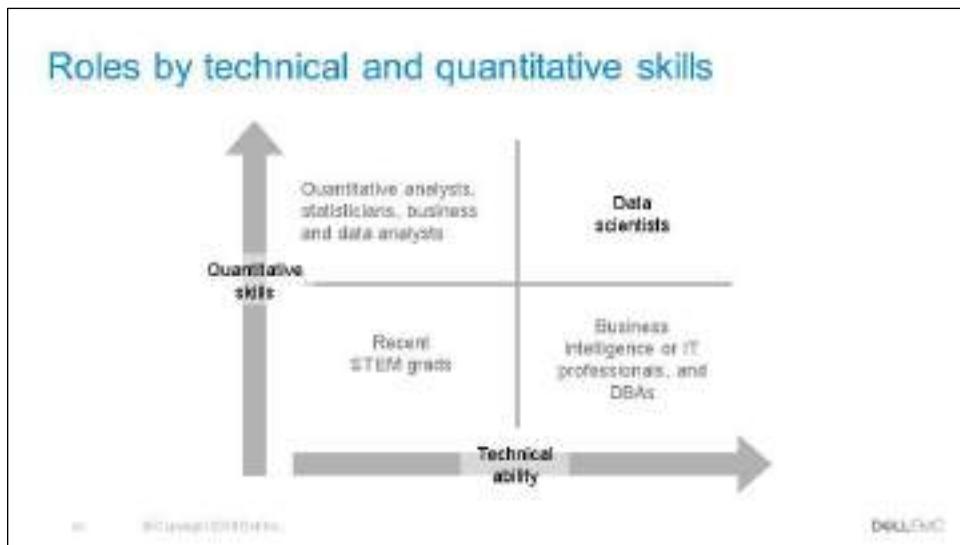
- Provision and administer analytical sandbox
- Manage large-scale data architecture
- Examples:
  - Database administrator
  - Programmer

DATA ENABLERS

The third category of people mentioned in the study is technology and data enablers. This group represents people providing technical expertise to support analytical projects, such as provisioning and administrating analytical sandboxes, and managing large-scale data architectures that enable widespread analytics within companies and other organizations. This role requires skills related to computer engineering, programming, and database administration.

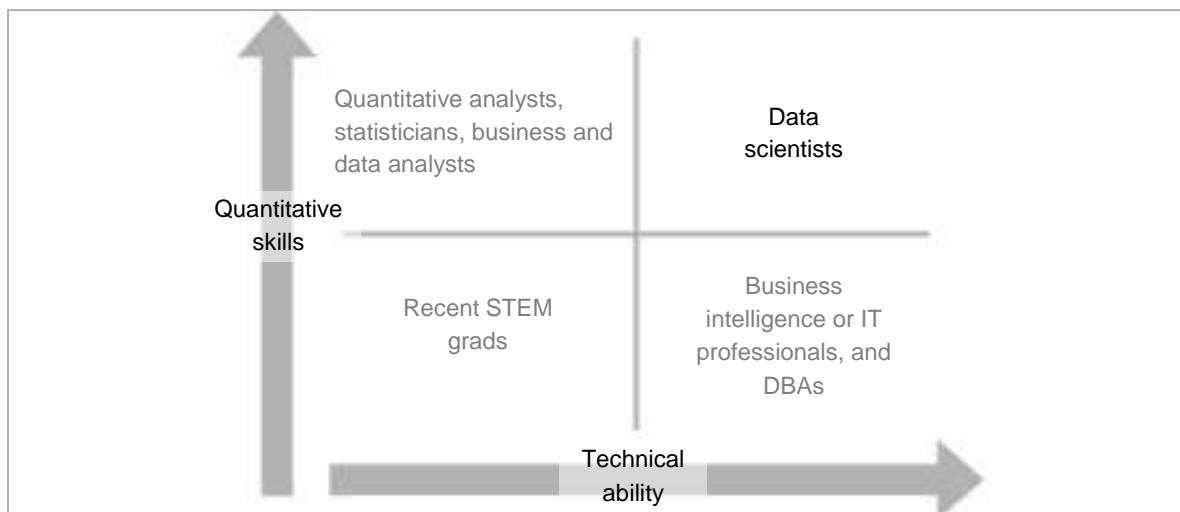
These three groups must work together closely to solve complex Big Data challenges. Most organizations are familiar with people in the latter two groups mentioned, but the first group, deep analytical talent, tends to be the newest role for most and the least understood. For simplicity, this discussion focuses on the emerging role of the data scientist. It describes the kinds of activities that role performs and provides a more detailed view of the skills needed to fulfill that role.

## Roles by technical and quantitative skills



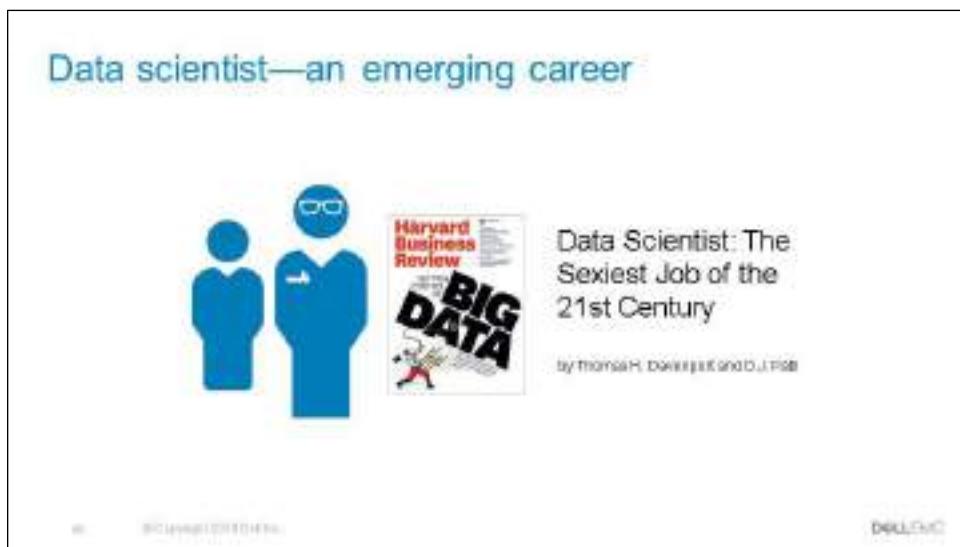
When looking for potential data scientists, you will find their technical and quantitative skills vary. Generally, they fall into the four categories:

- Recent science, technology, engineering, and mathematics (STEM) graduates
- Business intelligence or IT professionals, and DBAs
- Quantitative analysts, statisticians, business and data analysts
- Data scientists



Of these four areas, data scientists require the highest levels of both quantitative and technical skills.

## Data scientist—an emerging career



Because of the high value that organizations have placed on developing and retaining deep analytical talent, the data scientist has been dubbed the "sexiest" job of the 21st century. [hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/pr](http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/pr)

In 2016 and 2017, data scientist was the #1 Best Job in America according to Glassdoor: [www.glassdoor.com/List/Best-Jobs-in-America-LST\\_KQ0,20.htm](http://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm)

## Profile of a data scientist

**Profile of a data scientist**



**Quantitative skills**

- Expertise in mathematics and statistics

**Technical aptitude**

- Proficient programming skills
- Strong IT background

**Skeptical and critical thinking**

- Examine the work in a non-biased manner

**Curious and creative**

- Passionate about data
- Find novel ways to solve problems

**Communicative and collaborative**

- Articulate the business value in a clear way
- Collaboratively work with other groups

DATA SCIENCE

Data scientists are thought of as having five main sets of skills and behavioral characteristics.

Data scientists' skills and behavioral characteristics include:

- **Quantitative skills**, such as mathematics or statistics and expertise in quantitative skill.
  - **Technical aptitude**; namely, software engineering, machine learning, programming skills, and expertise in technical aptitude.
  - **Skeptical mindset and critical thinking**; it is important that data scientists can examine their work critically rather than in a one-sided way.
- 

## Lesson: Data scientist

- **Curiosity and creativity**, being passionate about data and finding creative ways to solve problems and portray information.
- **Communication and collaboration skills**; they must be able to articulate the business value in a clear way and collaboratively work with other groups, including project sponsors and key stakeholders.

Data scientists are comfortable using this blend of skills to acquire, manage, analyze, and visualize data and tell compelling stories about it.

## Check your knowledge

Check your knowledge

Which is an appropriate skill that a data scientist must have?

A. People management      C. Skeptical and critical thinking  
B. Financial management    D. Graphics design

Source: © 2018 Dell Inc.

DELL.COM

### Check your knowledge:

Which is an appropriate skill that a data scientist must have?

- A. People management
- B. Financial management
- C. Skeptical and critical thinking
- D. Graphics design

### Question:

Which is an appropriate skill that a data scientist must have?

### Answer:

---

## Module Summary

Key topics covered in this module were:

- Big Data and its characteristics
- Sources of Big Data
- Evolving analytical architecture
- The role of data scientist

10

DATA SCIENCE

DELL INC.

Key topics covered in this module were:

- Big Data and its characteristics
- Sources of Big Data
- Evolving analytical architecture
- The role of data scientist

# Data analytics lifecycle

## Introduction



### Data analytics lifecycle

Upon completing this module, you should be able to:

- ✓ List key phases of the data analytics lifecycle.
- ✓ Apply the data analytics lifecycle to a case study scenario.
- ✓ Describe the purpose of an analytics plan.
- ✓ Name the four core deliverables for a successful project.

This module provides an overview of the data analytics lifecycle, a framework to guide an analytics project team from the initial stages of a project through the project's successful completion.

## Lesson: Data analytics lifecycle overview

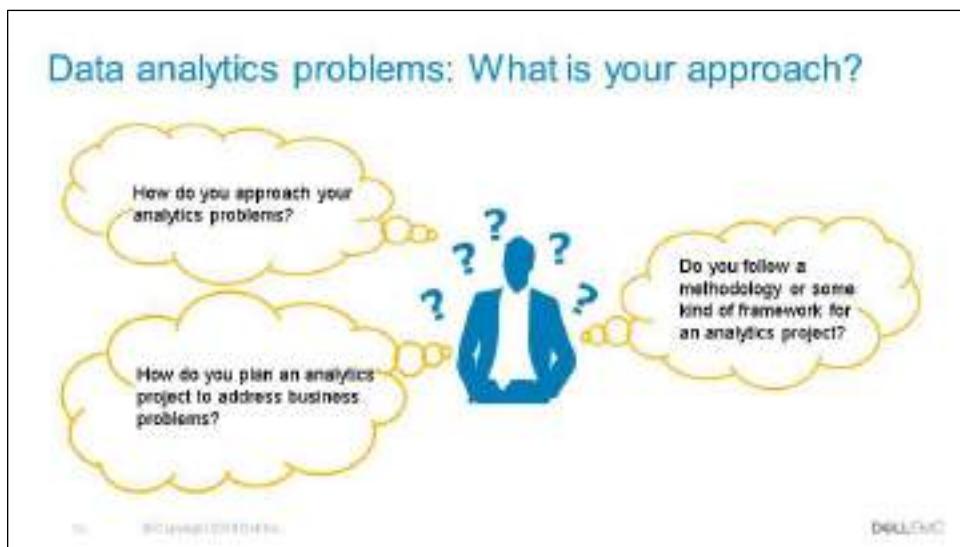
### Introduction

# Lesson: Data analytics lifecycle overview



This lesson provides an overview of the data analytics lifecycle.

## Data analytics problems: What is your approach?



Analytics focuses on developing new insights and understanding of business performance based on data and statistical methods. Organizations can apply their Big Data to uncover new, emerging trends, identify potential business opportunities, and discover new ways to gain competitive advantages. Consider how your organization currently addresses analytics problems and projects.



## Data analytics problems: What is your approach?



### Discussion

## Question/Discussion Topic: Data analytics problems: What is your approach?

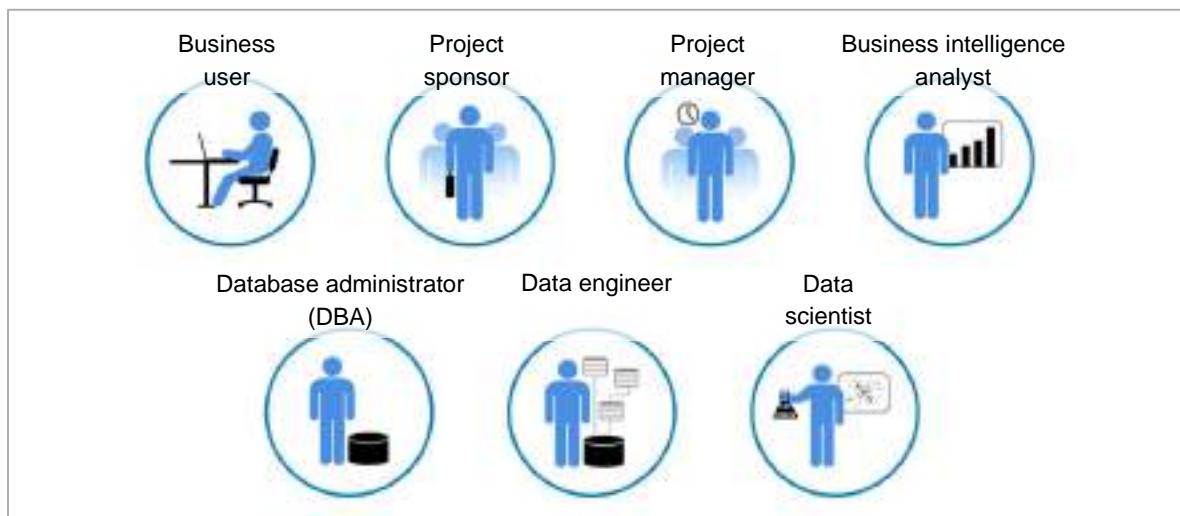
Data analytics problems: What is your approach?

### Discussion Notes:

## Key roles for a successful analytics project



This image depicts the various roles and key stakeholders of an analytics project. Each plays a critical part in a successful analytics project. Although seven roles are listed, fewer or more people can accomplish the work, depending on the scope of the project, the organizational structure, and the skills of the participants.



Here are the descriptions for each role:

- **Business user:** Someone who benefits from the end results and can advise the project team on the value of end results and how the project results will be operationalized.

## Lesson: Data analytics lifecycle overview

- **Project sponsor:** Responsible for the genesis of the project, providing the impetus for the project and core business problems. The project sponsor generally provides the funding and gauges the degree of value from the final outputs of the working team.
- **Project manager:** Ensures that key milestones and objectives are met on time and at the expected quality.
- **Business intelligence analyst:** Provides business-domain expertise with deep understanding of the data, KPIs, key metrics, and analytics from a reporting perspective.
- **Data engineer:** Applies deep technical skills to assist with data extraction from source systems and data ingestion on the analytic sandbox.
- **Database administrator (DBA):** Provisions and configures the database environment to support the analytical needs of the working team.
- **Data scientist:** Provides technical expertise for analytical techniques and data modeling, and applies the proper analytical techniques to given business problems to achieve the overall analytical objectives.

## Why use data analytic lifecycle

### Why use data analytic lifecycle

- Ensures the business problems are well-defined early in the project:
  - What is the desired business outcome?
  - How will success or failure be determined by the business stakeholders?
- Provides a comprehensive, repeatable method for conducting analyses
- Aids communicating key tasks and assignments within the team
- Plans and scopes the amount of work involved
- Properly sets expectations for the project stakeholders

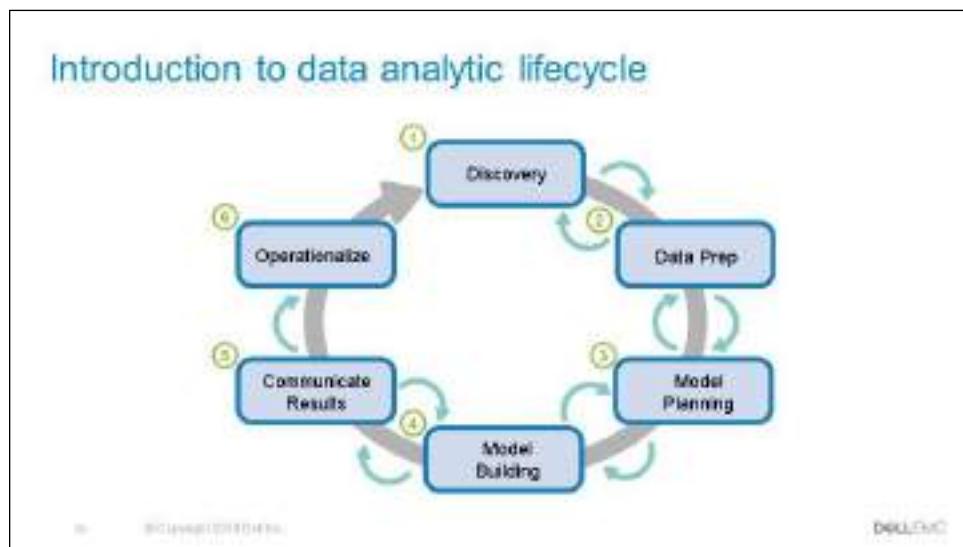
... [View full slide](#)

DELL.COM

Data science projects involve diagnosing and reframing business challenges as analytics problems and providing solutions to the problems. Many analytic projects seem huge and daunting at first, but a well-defined process enables you to break complex projects into smaller steps that can be more easily addressed.

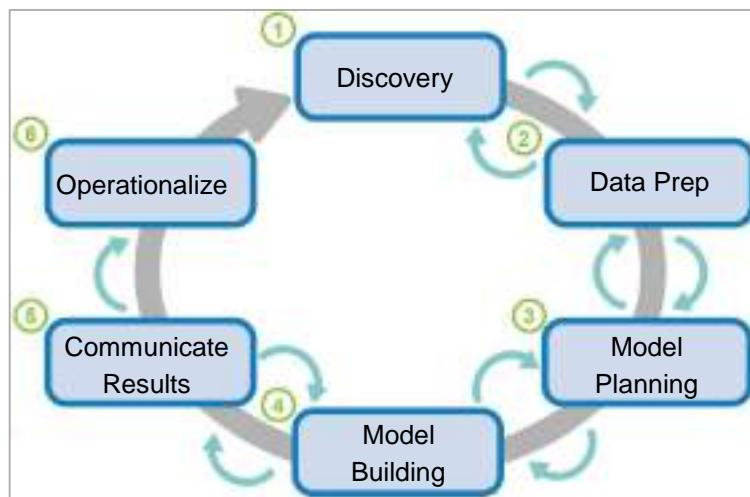
Having a well-defined process for performing analytics also ensures that you establish a comprehensive, repeatable method for conducting analysis. In addition, it helps focus the team's time early in the process on getting a clear grasp of the business problem to solve.

## Introduction to data analytic lifecycle



The data analytics lifecycle portrays a best practices approach for an end-to-end analytics process from discovery to project completion. There are six phases in the data analytics lifecycle. As new knowledge is obtained, movement from any phase to another and back again may occur throughout the lifecycle.

The circular arrows shown in the image are intended to convey that you can move iteratively between phases until you have sufficient information to continue. The callouts represent questions to gauge whether enough details are available or enough progress has been made to move to the next phase.



Here is a brief overview of the six phases of the Data Analytic Lifecycle::

- **Phase 1—Discovery:** Do I have enough information to draft an analytic plan and share for peer review?

- Learn the business domain, including relevant history, such as whether the organization or business unit has attempted similar projects in the past from which you can learn.
  - Assess the resources you must support in the project in terms of people, technology, time, and data.
  - Frame the business problem as an analytic challenge that can be addressed in subsequent phases.
  - Formulate initial assertions and questions to begin testing and exploring the identified datasets.
- At the end of the discovery phase, an analytic plan should be drafted. This plan maps out the major activities envisioned for the project and must be updated throughout the project as new datasets are identified or the modeling approach changes.
- **Phase 2—Data Preparation:** Do I have enough good quality data to start planning the model?
  - Prepare an analytic sandbox for the project.
  - Perform Extract, Load, and Transform (ELT) or Extract, Transform, and Load (ETL) activities in the sandbox.
  - Explore the data thoroughly, and take steps to condition the data.
- **Phase 3—Model Planning:** Do I have a good idea about the type of model to try?
  - Determine the methods, techniques, and workflow you intend to follow in the analysis.
  - Explore the data to learn about the relationships between variables. Then, select key variables and the models you are likely to use.
  - To prepare the data for the analysis, perform extra transformations and conditioning.
- **Phase 4—Model Building:** Is the model robust enough? Have we definitely failed?
  - Develop datasets for training and testing the model.
  - Fine-tune the code and sandbox to ensure efficient processing and repeatability of results.

## Lesson: Data analytics lifecycle overview

- **Phase 5—Communicate Results:** Does the expected benefit justify moving the model into production?
  - Determine if you succeeded or failed, based on the criteria established in the discovery phase, in collaboration with your stakeholders.
  - Identify your key findings, quantify the business value, and develop a narrative and presentations to convey your findings to stakeholders.
- **Phase 6—Operationalize:** Is the model performing as expected? Is any recalibration needed?
  - Deliver code and technical documentations.
  - Possibly run a pilot before fully implementing the models in a production environment.

## Lesson: Discovery phase

Discovery phase

Lesson: Discovery phase



## Discovery phase—key activities

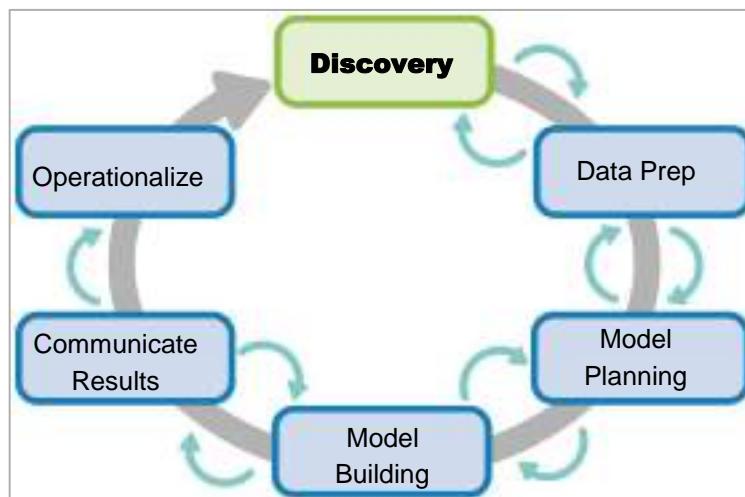
### Discovery phase—key activities

- Draft the business problem statement.
- Conduct stakeholder and business expert interviews.
  - Identify the current business state and any pain points.
  - Understand what related projects have been attempted in the past.
  - Refine the business problem statement.
- Reframe the business problem as an analytics challenge.
- Assess resource needs and availability.
- Draft an analytic plan.

© Copyright 2018 Dell Inc.

A well-executed Discovery phase can help ensure that the proper expectations for the project are set with all of the stakeholders and the project team members. After some early discussions with the project sponsor and a few business experts, it is important to document the business problem to be addressed.

Bearing in mind the draft business problem statement, your next step is to interview the various stakeholders and business experts. With the interview results, the problem statement is often refined or further clarified. Next, the business problem is expressed as an analytics challenge that will be addressed in the future phases. Finally, it is useful to draft an analytic plan that identifies the desired resources and the possible next steps in the project.



## Draft the business problem statement

### Draft the business problem statement



- Clearly articulate the current situation and pain points.
- Determine why addressing the problem matters.
- Include the intended outcome or ideal state.
- Tailor the statement to the key stakeholders.

Drafting the initial business problem statement should be accomplished with the project sponsor or the sponsor's staff, but other representatives may be included. A properly worded statement of the business situation and the intended outcomes will greatly help not only the entire project but also the next step of interviewing the various stakeholders and subject matter experts.

## Discovery—interviewing the project sponsor

**Discovery—interviewing the project sponsor**

Ask about:

- Business problem?
- Pain points?
- Expected outcome?
- If problem not addressed?
- Stakeholder obstacles?
- Constraints?
- Who could provide insights to problem and processes?
- Who has final approvals?



© 2018 Dell Inc. All rights reserved.

It is important to ask open-ended questions while conducting the interviews. Avoid simple yes/no questions such as, "Would you like to increase revenue?" It is important to listen to the project sponsor and understand any possible biases or predispositions.

For example, the project sponsor may focus on a possible solution and state something such as, "We need a Hadoop cluster to deploy a recommender system for the website." The interviewer should avoid arguing or getting into a solution tradeoff discussion. Rather, the interviewer should ask numerous follow-up questions that start with "why," such as "Why is that important?" or "Why is that a problem?"

Here are sample interview questions to use with the **Project Sponsor** in framing the business problem:

- What is the current business problem?
- What are the organization's current pain points?
- What are the expected business outcomes of the project?
  - How will the outcomes be measured?
  - Examples:
    - 15% increase in revenue

- 5% reduction in defects
- 10% increase in Net Promoter Score
- What will happen if the business problem is not addressed?
- What groups or stakeholders are likely obstacles?
- What are any constraints—for example, budget, resources, or timing?
- Who could provide further insights into the business problem?
- Who could provide further insights into the current business processes?
- Who has final approvals on the project's implementation?

## Discovery—interviewing other stakeholders and experts

### Discovery—interviewing other stakeholders and experts

Ask about:

- Your responsibilities?
- Pain points?
- Related project attempts?
- Existing tools/applications?
- Source systems and available datasets?
  - How to obtain access?
- Status of current IT infrastructure?
- Analytic sandboxes?
- Privacy concerns with datasets?
- Future IT infrastructure upgrades?
- Who else should be interviewed?
- Other insights or recommendations?

DATA SCIENCE

DELL EMC

The intent of these questions is to get a better understanding of the existing business processes and the technical resources, including datasets, that can be used for the project.

Here are sample interview questions to use with other **stakeholders and experts** in framing the problem and project constraints:

- What are your responsibilities?
- What are the pain points in the existing process?
- What related projects have been attempted in the past?
- What existing tools/applications are in use?
- What are the source systems and available datasets?
  - How do you obtain access?
- What is the status of current IT infrastructure?
- What analytic sandboxes are available?
- What privacy concerns are there with the datasets?
- Will the existing IT infrastructure be upgraded soon?
- Who else should be interviewed?
- What other insights or recommendations can you provide?

## Draft an analytics plan

### Draft an analytics plan

An **analytics plan** is the documentation used to guide the project team through the data analytics lifecycle. It is a living document that will be updated as we progress through different stages in the lifecycle.

Ask about:

- Provides a framework for understanding:
  - What has been accomplished so far
  - How the project should proceed
- Useful to remind the project team about the business objectives and the analytic approach
- As the analysis proceeds, analytic plan updates will often be necessary.
  - It is unlikely that everything will be known at the end of the Discovery phase
  - Scope changes will require additional communications with the project sponsor and stakeholders

...  
...  
...  
...

DATA SCIENCE

DELL INC.

Typically only one or two pages long, the analytics plan provides a high-level map of where the project has been and where it is going. The analytics plan's primary purpose is to help focus the data scientists and engineers on the analytical approach to the project.

## Analytics plan template

| Analytics plan template |       |   |
|-------------------------|-------|---|
| Component               | Phase | Description   |
| Business Problem        | 1     | A concise statement of the current business situation and why addressing the situation matters. |
| Business Impact         |       | The desired result of the project.  |
| Analytic Challenge      |       | The business problem expressed as an analytical problem.  |
| Data                    | 2     | Datasets and variables to be examined.  |
| Data Exploration        |       | Assessments about the data to be tested/validated.  |
| Proposed Model          | 3     | Analytic techniques to be applied and the model validation approach.                            |
| Key Findings            | 5     | New insights and the expected business benefits of implementing the project.                    |

Shown here is a template for an analytic plan mapped to several key phases of the data analytic lifecycle. The project team members should tailor the structure of the analytic plan to suit their specific needs. For example, if considerable effort is needed to stand up an analytics sandbox and perform ELT, then another row could be added to address that activity and key dependency.

## Sales analytics project—using data analytics lifecycle

### Sales analytics project—using data analytics lifecycle

**Scenario:** The Sales Operations team is struggling to provide the sales managers with useful insights and recommendations on the sales deals currently in the pipeline. The company's executives are frustrated with the constantly changing revenue estimates, particularly at the end of the quarter.

#### Business Objectives

The Sales Operations team is looking to accurately:

- Identify which sales deals are likely to be completed (booked) in the current quarter.
- Determine which sales deals are at risk.
- Estimate the quarterly revenue.

As the details of each data analytic lifecycle phase are presented, an actual analytics project will be used to illustrate the activities that occurred. In this example, the analytic projects involved better understanding the deals in the sales pipeline and identifying which deals were likely to be booked.

## Key discovery activities—sales analytics project

| Key discovery activities—sales analytics project  |  |
|---|--|
| <b>Interviews with Sponsor and Stakeholders to Understand Pain Points</b>   | <b>Business Problem Statement</b>  |
| <ul style="list-style-type: none"><li>• Perceived differences of sales deal forecast accuracy by geography and market verticals</li><li>• Good-looking deals are suddenly dropped or slip into the next quarter</li><li>• Lack of timely information and advanced warning to ensure deals are closed within the quarter</li></ul> | <ul style="list-style-type: none"><li>• The lack of timely visibility into which deals are likely to book within the current quarter is preventing the sales team from properly focusing on the correct deals and achieving the quarterly revenue goals.</li></ul> |

After you conduct several interviews with sales reps, sales operations personnel, sales managers, and executives, you see that the key point here is that it is enough to build a model to quantify the risk of each sales deal being booked in the current quarter. This model must be implemented as part of the current business processes.

## Key discovery activities—sales analytics project, cont.

| Key discovery activities—sales analytics project, cont  |  |
|---|--|
| <b>Analytic challenges</b> <ul style="list-style-type: none"><li>▪ Design a measure for quantifying the risk in each sales deal</li><li>▪ Enable the sales team with tools to:<ul style="list-style-type: none"><li>– Better inspect deal risk</li><li>– Help move deals through the sales pipeline</li><li>– Convert high-risk deals to low-risk deals</li></ul></li></ul> | <b>Data sources</b> <ul style="list-style-type: none"><li>▪ Key data source – Customer Relationship Management (CRM) system to track the sales opportunities</li><li>▪ Other identified data sources provided:<ul style="list-style-type: none"><li>– Customer demographics</li><li>– Previous product purchases and installations, product service history</li><li>– Service contract renewals</li><li>– Sales representative details (Tenure, Sales Manager, Region)</li></ul></li></ul> |

Through the Discovery process, it was decided to merge the sales deals in the CRM system with other datasets related to the customer demographics, previous customer purchases and product installations, service history, contract renewals, and characteristics about the sales representatives.

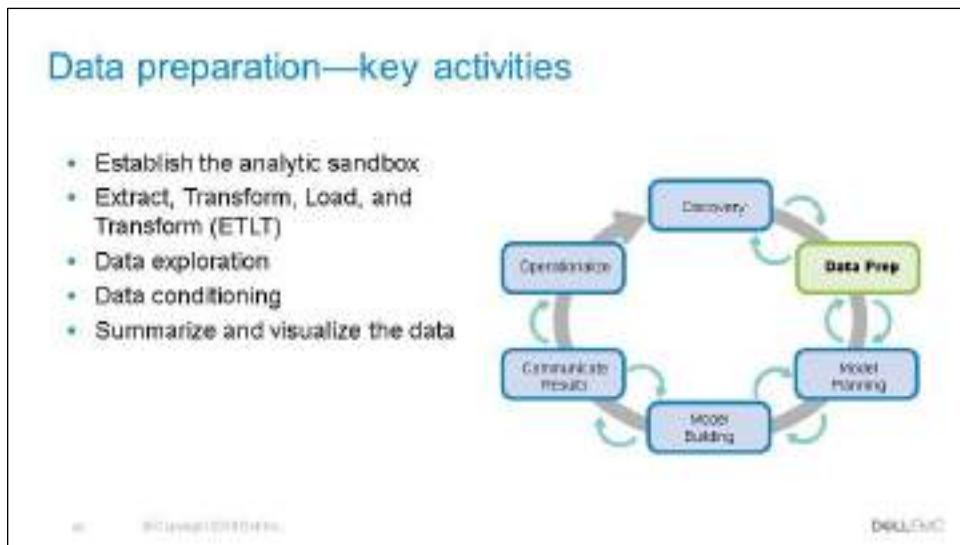
## Lesson: Data preparation phase

Data preparation phase

Lesson: Data preparation  
phase

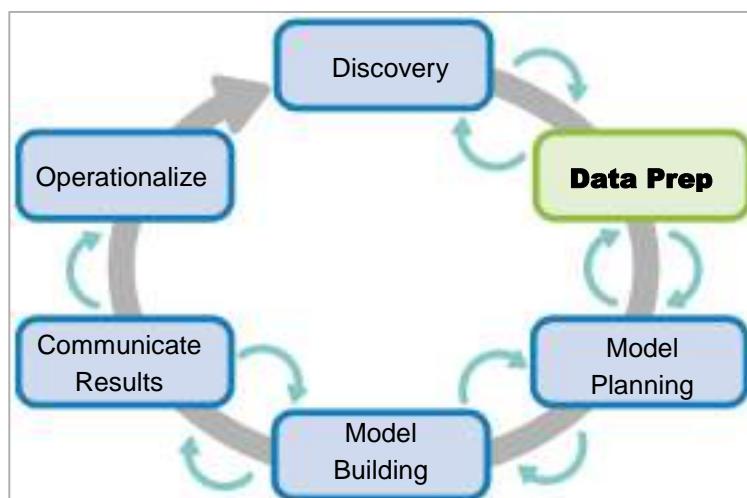


## Data preparation—key activities



In the Data Preparation phase, the project team:

- Loads data into the analytic sandbox
- Understands the datasets in great detail
- Cleans and conditions the data
- Begins preparing the data for the model planning and building phases



## Establish the analytic sandbox

### Establish the analytic sandbox

- Estimate the size of the datasets to be loaded.
- Plan for a sandbox about 5 to 10 times the size of the original datasets.
- Consider the data and analytic tools to be utilized.
- Address privacy concerns and security measures.
- Work with IT to prepare the sandbox.



DELL EMC

One of the first activities is to identify a location to perform the data exploration and eventual model building tasks. For very large datasets, it is typical to prepare a dedicated analytic sandbox area in which the project team can work. A dedicated sandbox is often preferred to minimize any impact to production systems for transaction processing or reporting. To allow space for data manipulation and experimentation, the analytic sandbox should be 5 to 10 times the size of the data load.



The type of sandbox often depends on the structure of the data to be examined. If the data is highly structured, a relational database may suffice. If the data is highly unstructured, the analytic sandbox may be a Hadoop cluster.

It is important to engage IT to establish the sandbox so that it is owned by analysts, but also to ensure that any privacy and security concerns are

addressed. The use of a suitable sandbox helps ensure that copies of the data are not loaded on individual workstations and other devices.

## Extract, transform, load, and transform (ETLT)

### Extract, transform, load, and transform (ETLT)

- Ideally, extract and load the data as stored within the source system.
- Ensure the list of variables included in the sandbox is exhaustive.
- Transform the data within the sandbox.



After a sandbox is established, it is useful to load the data in its original format, or at least with minimal transformations applied to the data. Oftentimes, the project team requires assistance from the other organizations to extract and load the data into the sandbox. If any transformations are performed before loading into the sandbox, the project team will not necessarily have any visibility into the raw, unfiltered data, which may lead to unintended and misleading results.

For example, if transactions from an online shopping website are loaded into the sandbox, it is possible that the data will only include the items that were in the shopping cart at the time of checkout. Items added but removed from the shopping cart may be of utmost importance to the analysis. Furthermore, if the data must be reloaded into the sandbox but the transformations were incorrect, the team must wait for the data to be extracted again. So, the recommended approach is ELT.



## Data exploration

**Data exploration**



- Work with business experts to understand what the data elements mean.
- Identify missing data entries.
- Understand how to join datasets or extract the desired information.
- Identify outliers and possible data quality issues.

© 2014 DELL INC. DELL.COM

A data Scientist often plays the role of a detective to understand what the various data elements and values represent.

The data scientist and the project team should leverage business experts who are familiar with the datasets and the business processes used to create the data. The data scientist will be skeptical and validate any assertions the experts make.



## Data conditioning

### Data conditioning



- Join and merge the datasets.
- Cleanse the data.
- Normalize datasets.

DATAVIZ

Data conditioning refers to the process of cleaning data, normalizing datasets, and performing transformations on the data in preparation for further analysis. This step deals with detecting and removing errors and inconsistencies from data to improve its quality.

## Summarize and visualize datasets

### Summarize and visualize datasets



- Prepare summary statistics and visualizations to better understand the data.
- Identify possible variables that may be related to the outcome to be modeled.

If the data conditioning step is performed by IT, the data owners, a DBA, or a data engineer, it is also important to involve the data scientist, because many data conditioning decisions can affect subsequent analyses.

For example, it may be decided to filter out apparent duplicate transactions, but these duplicates may be of particular value to the analysis. It would be better to flag the suspect records and let the data scientist address the inclusion of such records as appropriate.



## Key data preparation activities—sales analytics project

Key data preparation activities—sales analytics project

A Greenplum database was selected as the analytic sandbox:

- Some datasets were already loaded into Greenplum.
- Most of the data was already in another SQL database or well structured.

Data conditioning challenges that needed to be addressed:

- No historical data on the sales deals were maintained.
- Some deals were in parent/child relationships:
  - For example, a deal would ship products to multiple countries or U.S. states.
- Completeness of CRM data was somewhat salesrep dependent:
  - For example, supplement missing customer data with customer data from other sources.

DATA SOURCE: DELL INC.

The project team used a Greenplum database to store and analyze the various datasets. Based on PostgreSQL, a Greenplum database is a shared-nothing, massively parallel processing design that emphasizes efficiency and linear scalability.

The team ran into several data issues that needed to be addressed for the project to continue. For example, the available CRM data only provided the current state of the deals. However, to build a predictive model, it would be necessary to reconstruct the data to show what was in the CRM system one, three, or 12 months ago.

In addition, since the CRM data was dependent on the diligence of the sales representative to enter accurate and complete data, it was sometimes necessary to supplement the CRM data elements with data from other sources.

## Key data preparation activities—attributes

### Key data preparation activities—attributes



Data sources:

- Sales opportunities
- Install base
- Service history
- Purchase history
- Sales representative details
- Customer demographics
- Service contracts

DATA SOURCE: DELL INC.

These attributes are some examples of the data elements to consider, in building the model.

### Sales opportunities

- Opportunity number
- Account name
- Company location
- Opportunity create date
- Opportunity close date
- Number of products
- Number of changes/updates
- Forecast amount
- Stage—opportunity status
- Quote attached
- Competitor detail
- Business solution detail
- Partner detail

### Install base

- Number of installed/de-installed products
- Average age of assets



### Service history

- Total ticket time
- Total logged labor hours
- Number of service tickets
- Severity level

### Purchase history

- Number of product families
- Number of purchases—annual/quarter
- Number of upgrades
- Product type/family

### Sales representative attributes

- Rep tenure
- Rep tenure in role
- Historical conversion rates
- Average deal size
- Percentage of goal attainment

### Service contracts

- Renewals history

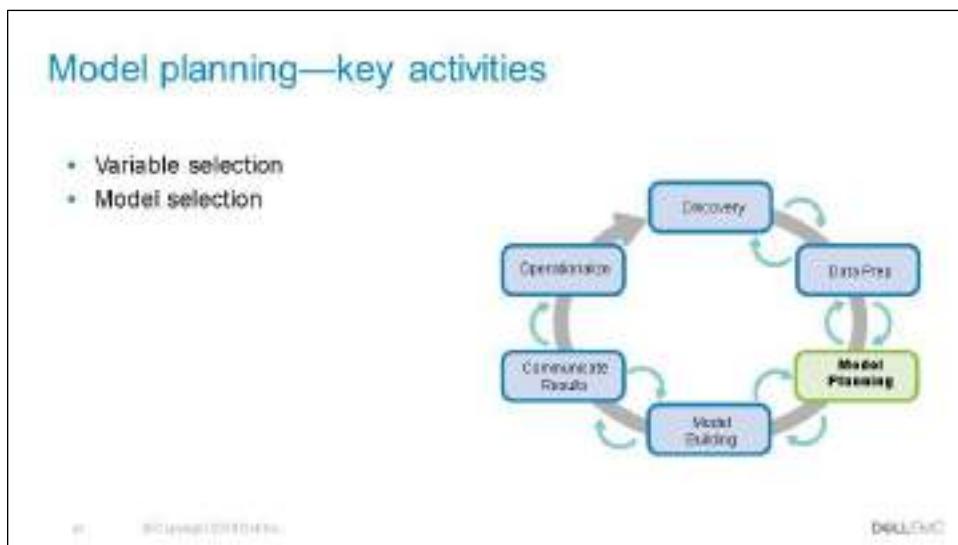
## Lesson: Model planning phase

Model planning phase

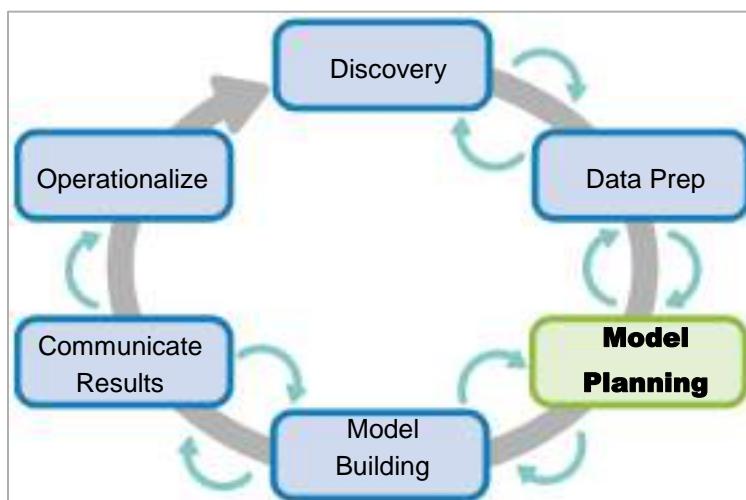
Lesson: Model planning  
phase



## Model planning—key activities



Based on the analytics challenge, the project team selects a possible analytical model and the corresponding variables and other inputs.



This model-planning phase may seem straightforward, but it is often necessary to perform extra data exploration, data conditioning, and transformations to prepare the data for the model building phase.

## Variable selection

### Variable selection

- Explore the data and understand the relationships among the variables.
- Question and evaluate opinions from stakeholders and domain experts.
- Identify relationships or correlations:
  - Among possible input variables
  - Between input and outcome variables
- Leverage a technique for dimensionality reduction, if applicable.
- Perform additional data transformations to prepare variables for modeling.

40

DATA SCIENCE

DELL INC.

In this phase, the main objective of the data exploration is to understand the relationships among the variables. Not only is it important to understand the relationships between the input variables and the outcome variables, it is important to understand the relationships, if any, between the input variables. When there is significant correlation between two or more input variables, it may be useful to perform some variable reduction activities.

For example, in healthcare-related analyses, the height and weight of patients is often highly correlated. So, instead of using both height and weight individually, it may make sense to use the Body Mass Index (BMI) as a predictor. BMI is a person's weight in kilograms divided by the person's height in centimeters squared.

Also, there are techniques such as Principal Component Analysis (PCA) to reduce the dimensionality of the data. In this phase, all candidate variables should be considered; don't reduce 100 variables to what the team thinks are the four most important variables. The most useful variables will be chosen in the model-building phase.

## Model selection

### Model selection

- Choose an analytical technique or a shortlist of candidates based on:
  - The purpose of the analysis (for example, exploratory or prediction)
  - The types of input and outcome variables (for example, categorical or continuous)
- Decide to fit one model or a series of models:
  - For example, one regression model to handle 50 U.S. states, or 50 regression models—one model for each US state
- Determine the analytic tool to fit the selected model.

10

DATA SCIENCE

DELL INC.

The main goal is to choose an analytical technique or a shortlist of candidates based on the end goal of the project. It is often useful to revisit the analytic challenge at this stage of the project, to ensure that the analytic challenge is still relevant and that there is not any scope creep in the project.

From the context of this course, a model is discussed in general terms to refer to the analytical techniques to be applied. One observes events happening in a real-world situation and attempts to construct models that emulate this behavior. For machine learning and data mining, these rules and conditions are grouped into several general sets of techniques such as classification, association rules, and clustering.

Further, in this phase, the suitability of existing analytic tools should be examined. It may be necessary to add another tool to the analytic sandbox.

## Key model planning activities—sales analytics project

### Key model planning activities—sales analytics project

The team decides to classify each sales opportunity into one of three categories:

- Book—the sales deal will be booked in the current quarter.
  - Push—the sales deal will be deferred into the next quarter.
  - Close—the existing opportunity will not result in a booking in any quarter.
- A multinomial logistic regression model was selected:
- The model output provides a probability for each of the three categories.
  - The opportunities can then be ranked to determine priorities for the sales team.

The project team selected multinomial logistic regression to model the sales data, for the following reasons:

- This determines the probability of one the three possible events—Book, Push, or Close—to occur.
- The probabilities could then be used to rank each opportunity for each sales manager or sales region.
- These rankings could be used to set priorities for the sales to address.

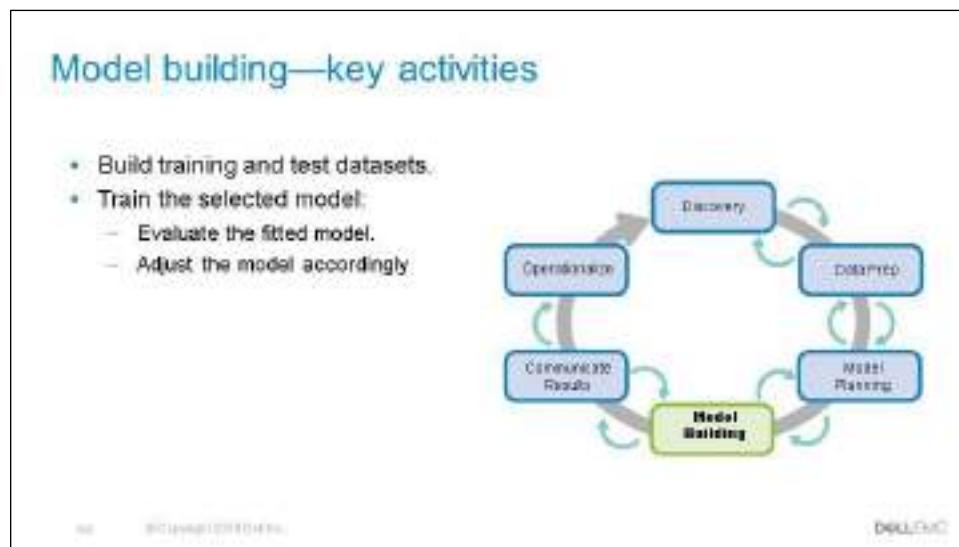
## Lesson: Model Building Phase

Model Building Phase

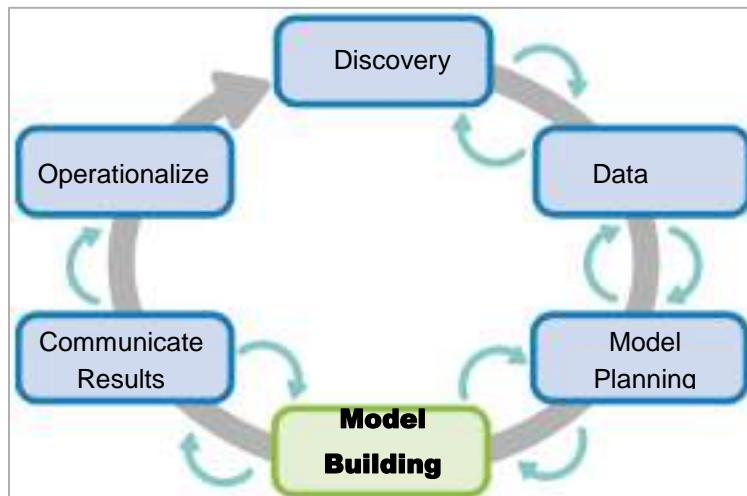
Lesson: Model Building  
Phase



## Model building—key activities



In the model-building phase, the selected analytical technique is applied to a set of training data. This process is known as "training the model." A separate set of data, known as the testing data, is then used to evaluate how well the model performs.



Often, the fitted model is to be applied to future observations. So, it is not typically sufficient to obtain the best model that explains all of the data; you must build a model that adequately predicts future events.

## Build training and test datasets

### Build training and test datasets

A typical approach is as follows:

- Randomly select a subset (say 70–90%) of the available data to train the model (**training dataset**).
- Use the remaining data as the **testing dataset** to evaluate the performance of the fitted model.

Additional considerations:

- Apply stratified random sampling to ensure proper representation from all groups.
- Does the modeling technique implicitly account for the training and testing datasets?

001

DATA SCIENCE

DELL INC.

Often, the building of the training dataset and the testing dataset is fairly straightforward:

- Using a proper random number generator, randomly assign a number, 0 through 1, to each record in the dataset.
- Select the records corresponding to the smallest X%—for example, 80%—of the assigned numbers; this dataset is the training dataset.
- The remaining ~20% of the records is the testing dataset.

The importance of random sampling is that the entire dataset may be already sequenced by attributes such as customer, geography, or time. So, using the first 80% of the records would bias the resulting model and exclude key categories from the training and testing datasets.

Sometimes, it may be necessary to deploy a method such as stratified random sampling to ensure adequate representation of all the key groups—or, strata—within the dataset. For example, if data from 50 U.S. states is to be analyzed, it may be very important to ensure that each state is adequately represented in the training and testing datasets.

In some more advanced analytical modeling techniques, such as Random Forests, the technique itself randomly extracts the data, fits the model, and tests the model. This process is then repeated several times.

## Train selected model

### Train selected model

- Identify the useful input variables (feature selection).
- Avoid overfitting the data.
- Verify model assumptions.



After the training and testing datasets are built, the model training can begin. In general, the purpose is to find the most useful input variables. This process is known as *feature selection*. This process is often dependent upon the particular modeling technique, but usually adding more variables improves the fit of the model.

Adding more variables will seldom make the fit worse, but it may make the predictive power of the model worse. As a warning against overfitting a model to a dataset, John von Neumann declared, "With four parameters, I can fit an elephant, and with five I can make him wiggle his trunk."

## Evaluate model

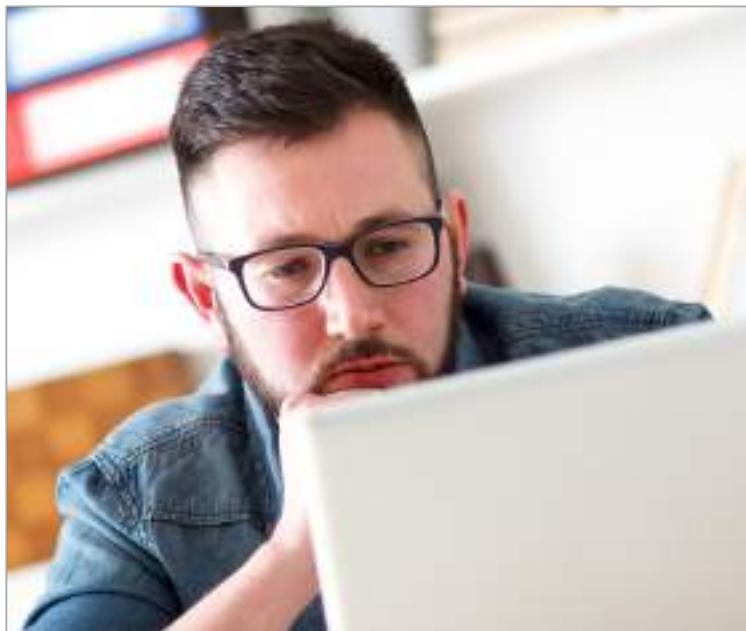
### Evaluate model

- Is the model accurate enough to meet the goal?
- Are there edge cases that are not properly handled?
- Does the model output/behavior make sense to the domain experts?



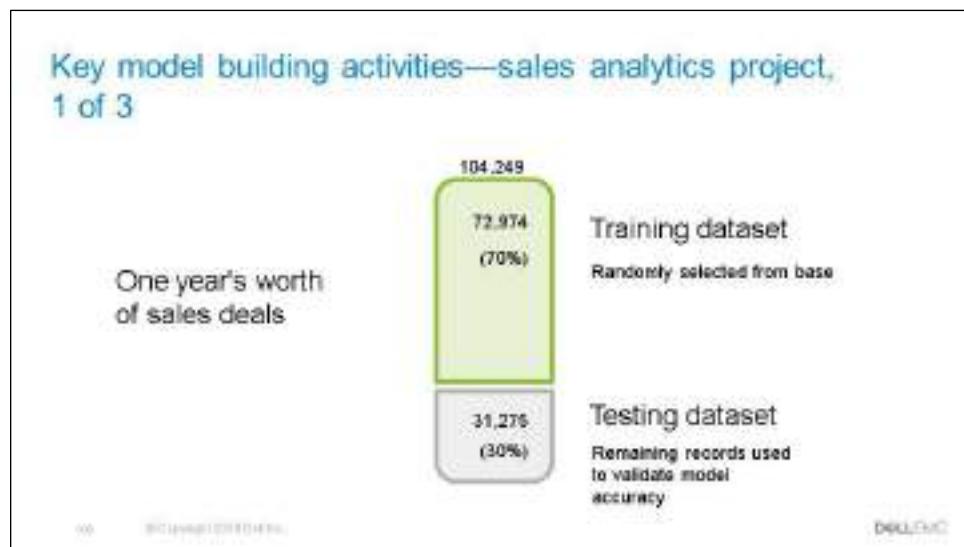
DATA SCIENCE

After the best input variables are selected, the next step is to ensure that the model assumptions are verified. For example, in linear regression the error terms are assumed to be normally distributed, with a constant variance. If the modeling assumptions are not met, it may be necessary to further transform the variables, include more variables, or modify the analytic approach.

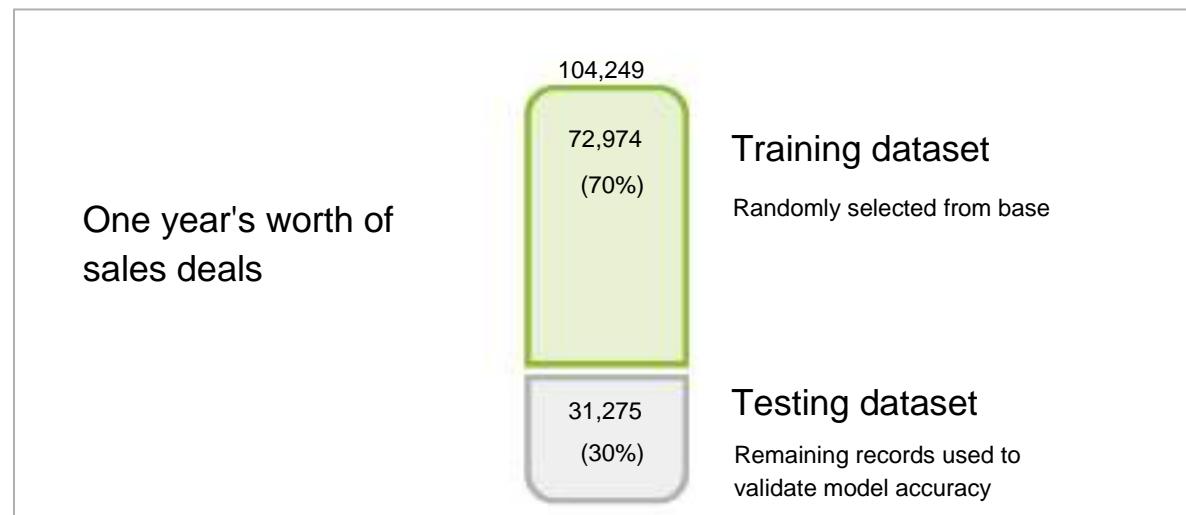


After a model is deemed acceptable, based on the training set, the next step is to validate the model against the testing set. If the desired predictive power is not met by the model, go back and rebuild it, but be careful not to bias the new fitted model to adequately explain the testing dataset.

## Key model building activities—sales analytics project, 1 of 3



From one year's worth of sales deals, 70% of the records were randomly selected to train the model, and the remaining 30% were used to validate the model.



## Key model building activities—sales analytics project, 2 of 3

### Key model building activities—sales analytics project, 2 of 3

Initial model built and validated:

- The multinomial logistic model was trained and key features (input variables) selected.
- The model results were shared with the domain experts.

Based on domain experts' feedback:

- The direct sales team will create price quotes well in advance of a deal booking.
- Deals from other teams will not add a quote to the system unless the deal is ready to be booked.
- Thus, attaching quotes is not a reliable predictor for deals through the channel partners.

Sales data stratified into three groups – direct sales, indirect sales, other sales:

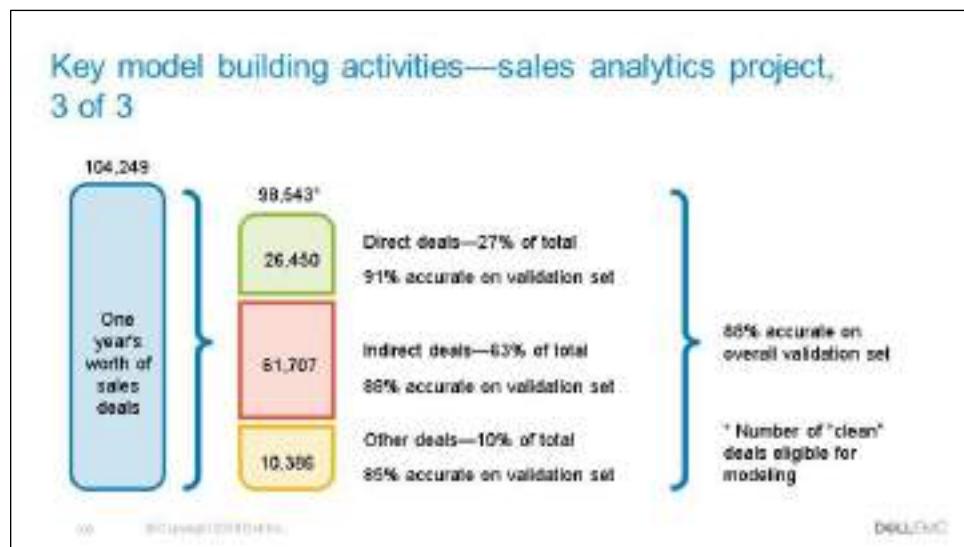
- Distinct training and testing datasets were built for each group.
- Three separate multinomial logistic models were built for each group.

400 800 1200 1600 2000

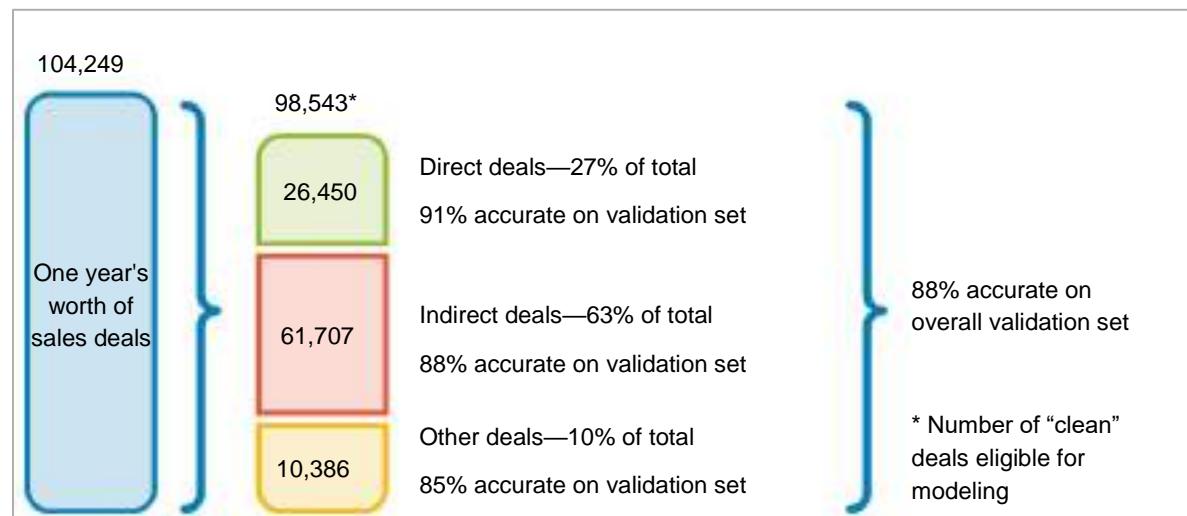
Dell EMC

After a multinomial logistic model was fit to the data, and the project team was satisfied with the results, the model results were shared with several domain experts. Quickly, it became apparent that several input variables and the timing of their addition to the CRM system depended on which sales group was managing the deal. So, it was decided to build a separate model for each sales group.

## Key model building activities—sales analytics project, 3 of 3



To split the data into the three different groups, it was necessary to exclude about 6,000 deals from the analysis. The legacy data was not clean enough to properly categorize the deals. Training and testing datasets were built for each group. Each group's model was trained and validated with an overall accuracy of 88%.



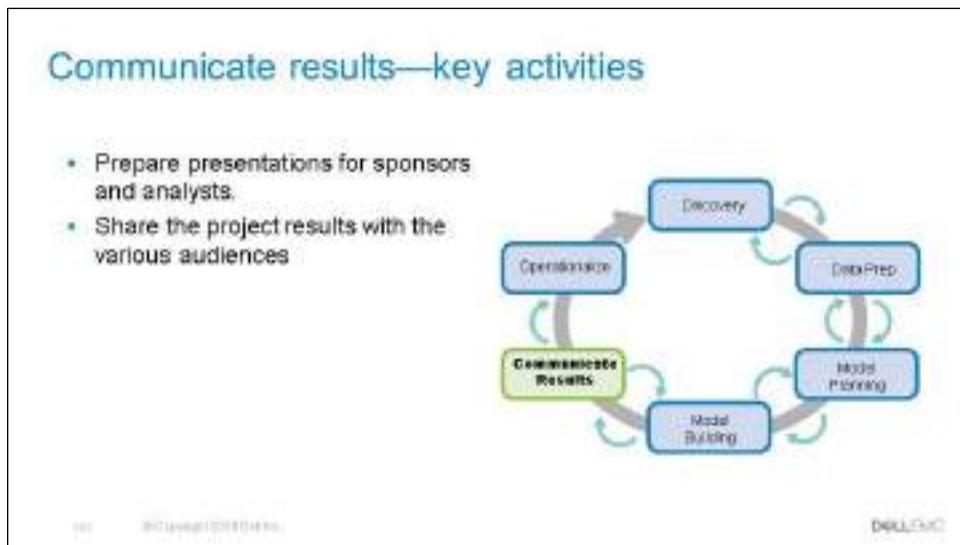
## Lesson: Communicate results phase

Communicate results phase

Lesson: Communicate  
results phase



## Communicate results—key activities



After an acceptable model is obtained, the next step is to communicate to the project sponsor and the stakeholders the project's findings and the business value of the model. If the desired business outcome was not obtained, this result also must be communicated.

## Prepare presentation for sponsors and analysts

### Prepare presentation for sponsors and analysts



- State the business problem and the objective of the analytic project.
- Provide the business value of implementing the model.
- Prepare recommended next steps.

While the Communicate Results phase may seem to be the simplest and most straightforward phase, it can often be the most challenging part of the exercise. The team must not only provide the outcome of their efforts but, in some circumstances, must also overcome the stakeholders' preconceived notions of what the most influential input variables are.

Furthermore, unless the project sponsors and stakeholders are kept periodically updated on the project as well as the objective of the project, it is typical for someone to state, "This is not what I asked for!" So, it is important to remind the audience about the business problem and the scope of the project.



## Share project results with various audiences

### Share project results with various audiences



- Build a strategy to communicate the findings
- Present the findings to the project sponsor and stakeholders

Consider how best to articulate the findings and outcome to the various stakeholders. Also, consider and include caveats, assumptions, and any limitations of results. Remember that, many times, the presentation is circulated within the organization, so be thoughtful about how you position the findings, and clearly articulate the outcomes.

Make recommendations for future work or improvements to existing processes. Also, this is the phase where you can underscore the business benefits of the work and begin making the case to eventually put the logic into a live production environment.



## Lesson: Communicate results phase

The project sponsor and business stakeholders will likely be interested in the business benefit of the project, not the technical details of how the model was built and refined. Often, it is beneficial to obtain buy-in from the rest of the stakeholders before sharing the results with the project sponsor. So, it is useful to have a strategy in place to build consensus on the project's results, to obtain approvals from the various stakeholders to proceed to the operationalize phase.

## Core deliverables to meet stakeholders needs

### Core deliverables to meet stakeholders needs

- Presentation for project sponsors and other executives:
  - "Big picture" takeaways for executive-level stakeholders.
  - Determine key messages to aid their decision-making process.
  - Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp.
  - **Objective:** Demonstrate the business benefit of implementing the model.
- Presentation for analysts:
  - Business process changes
  - Reporting changes
  - Fellow data scientists will want details and technical graphs (for example, ROC curves)
  - **Objective:** Demonstrate the validity of the model.
- Code
- Technical specifications for implementing the code

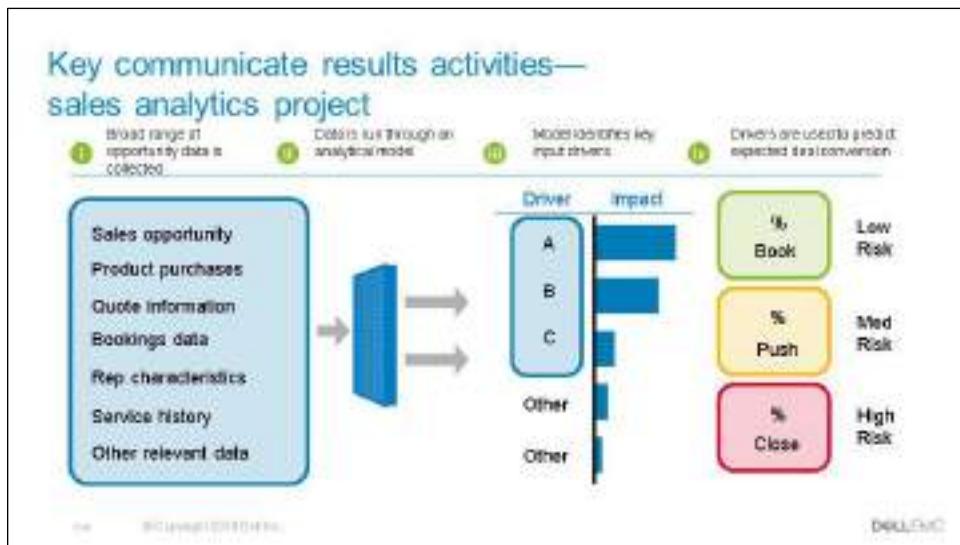
100 | BIG DATA ANALYTICS v2

DELL EMC

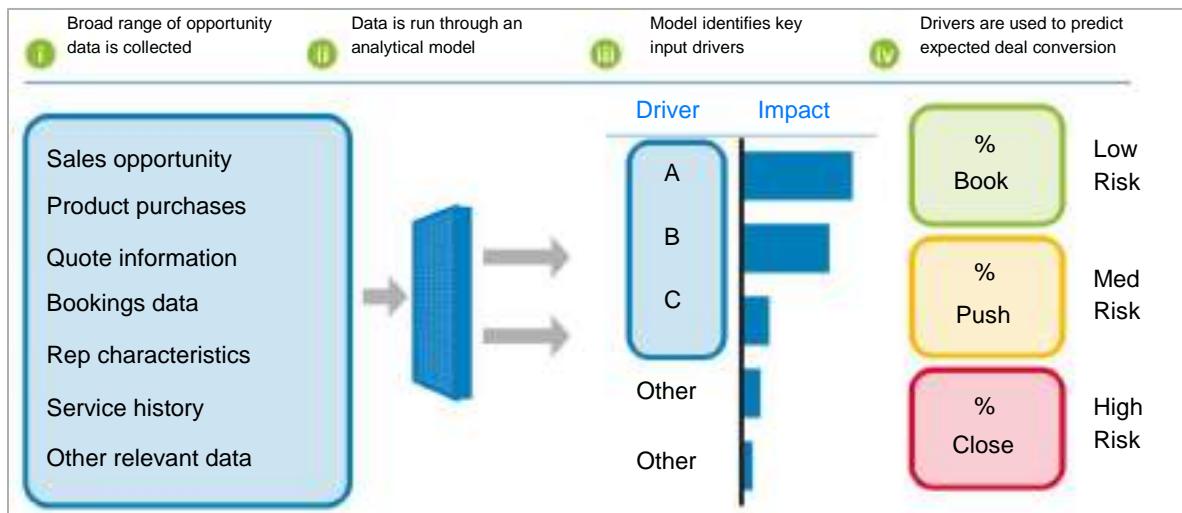
Although there may be several key stakeholders of an analytics project, such as executives, data engineers, and business users, most of their roles and responsibilities usually overlap at this stage of the project and can be met with four main deliverables.

The two presentations are instrumental in obtaining the business buy-in and approval to proceed with the Operationalize phase. Two sets of technical documentation, code and specifications, are necessary to begin the Operationalize phase. These four items are key deliverables in any successful analytic project.

## Key communicate results activities—sales analytics project



This is an example of how the modeling process could be communicated to the various stakeholders without going into great technical detail. Since three separate multinomial logistic regression models were built, each model will likely have its own set of key input drivers. How many of these details you share will depend on the audience to which you are presenting.



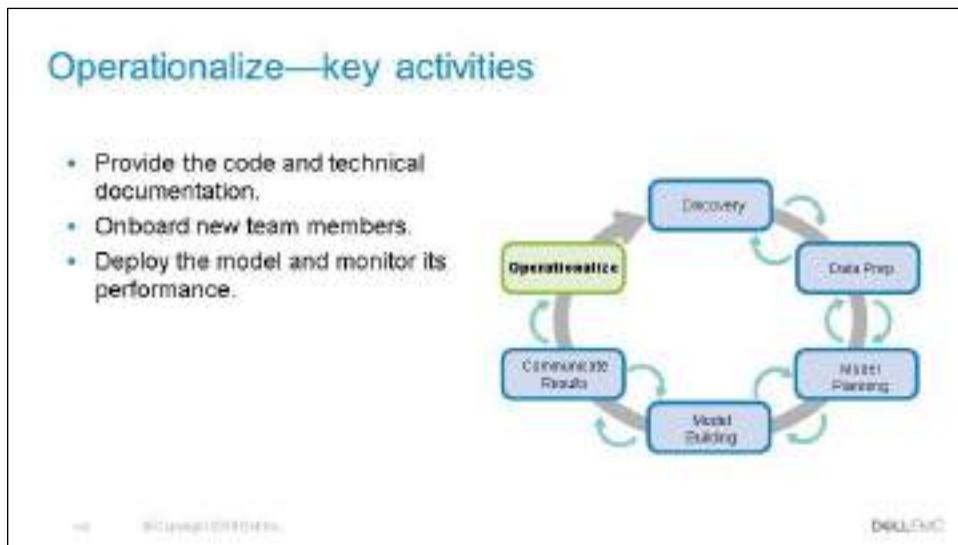
## Lesson: Operationalize phase

Operationalize phase

Lesson: Operationalize  
phase



## Operationalize—key activities



After the stakeholders have agreed to implement the model in the production environment, the Operationalize phase begins. Depending on the organization, the project team may be responsible for the model's implementation or may simply transfer the code and other technical documentation to a different team. In either case, it is important to bring any new participants up to speed on the model and the project's findings. During this phase, it is also important to establish the approach to monitor the performance of the model after it is placed into production.

## Provide code and technical documentation

### Provide code and technical documentation



- Complete documentation.
- Obtain acceptance from production owners.

Although the documentation might have been completed in the Communicate Results phase, it is important that it be reviewed with the production owners and other technical individuals who will implement and maintain the model. Try to use a collaborative approach so that everyone owns this documentation and it is not simply a help-desk ticket that is being submitted. It is important to obtain buy-in from the new team members about the business value of the project.

## Onboard new team members

Onboard new team members



- Review history of the project.
- Walk through the analyst's presentation.
- Demonstrate the business value.

148 802 Unpublished Material

DELL EMC

Most of the material developed in the sponsor and analyst presentations can be reused to onboard any new team members. Attempt to keep these individual concerns in mind. For example, team members responsible for the production environment will be concerned about performance of the model and any possible negative impacts to the end user.

## Deploy model and monitor

### Deploy model and monitor

- Deploy the model in test environment.
- Run a pilot project in production.
- Determine some monitoring of the implemented model.
- Recalibrate the model based on feedback.



Implementing the model in a test environment helps minimize any impacts to production. It is common to run a pilot program before fully implementing the model in production. Running a pilot helps minimize risk and further demonstrate the business value. Consider running the model in a product environment for a discrete set of single products or a single line of business, which will test your model in a live setting. This process allows you to learn from the deployment and make any needed adjustments before launching across the enterprise.

After the model is placed into production, it is often necessary to monitor the model's performance and establish a process to retrain and update the model. It should be noted that further communication of results often occurs during the Operationalize phase; the executives will want to know the actual ROI from their investment.



## Key operations activities—sales analytics project

### Key operations activities—sales analytics project

Using the developed models:

- A process was put in place to update the deal predictions weekly.
- The results were packaged into an easily consumable set of documentation.
- The appropriate content was shared with the sales reps, sales managers, regional directors, and so on.

The model was implemented into a process to provide weekly updates on the likelihood of a deal being booked, pushed, or closed. The project results presented here were just some of the initial successes the sales organization realized.

## Lesson: Conclusion

Conclusion

Lesson: Conclusion



## Lifecycle continuation

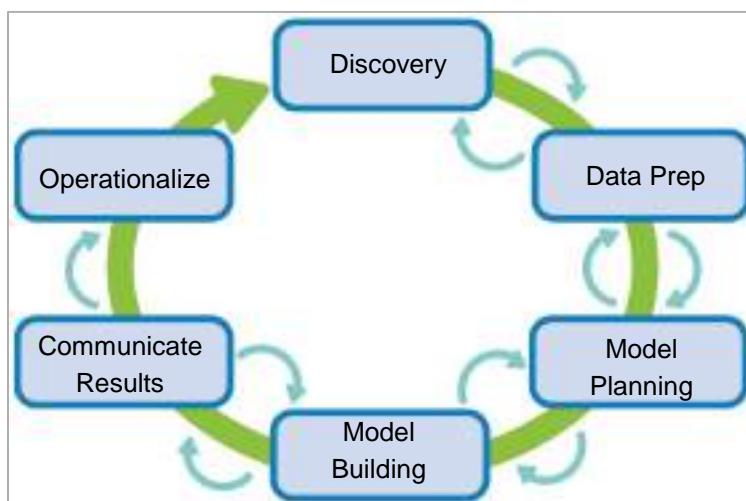
### Lifecycle continuation

The work does not end with the operationalize phase:

- The model needs to be monitored and possibly further refined.
- New attributes can be considered for inclusion in the models.
- The delivery mechanism can be simplified with self-service reporting.

DELL EMC

As the lifecycle graphic illustrates, the work does not end with the operationalize phase; there is often a continuous state of discovery and further refinement of the model or expansion to address other business problems such as estimating the quarterly revenue.



As the project team learns how the model results are being used, the delivery mechanism can be simplified with self-service reporting.

## Concepts in practice—Pivotal Greenplum Database

Concepts in practice—Pivotal Greenplum Database



**GREENPLUM.**

- Greenplum Database is an advanced data warehouse.
- Shared-nothing, massively parallel processing design emphasizes parallelism, efficiency, and linear scalability.
- Provides rapid analytics on petabyte-scale data volumes.
- Greenplum Database is based on PostgreSQL.
- Greenplum Database includes features designed to optimize PostgreSQL for analytics workloads.

100% GREENPLUM. 100% DELL.

Greenplum Database is an advanced, full-featured, open-source data warehouse. The shared-nothing, massively parallel processing architecture used by Greenplum incorporates the highest level of parallelism and efficiency to handle complex business intelligence and analytical processing.

It provides powerful and rapid analytics on petabyte-scale data volumes. Uniquely geared toward Big Data analytics, Greenplum Database is powered by the world's most advanced cost-based query optimizer, delivering high analytical query performance on large data volumes.

Greenplum uses this high-performance system architecture to distribute the load of multi-terabyte data warehouses, and can use all of a system's resources in parallel to process a query. Greenplum Database is based on PostgreSQL open-source technology. It is essentially several PostgreSQL database instances acting together as one cohesive Database Management System (DBMS).

Greenplum Database also includes features designed to optimize PostgreSQL for analytics workloads. For example, Greenplum has added parallel data loading—external tables—resource management, query optimizations, and storage enhancements that are not found in regular PostgreSQL.



Various power analytic tools are available for use with Greenplum Database:

- MADlib, an open-source, MPP implementation of many analytic algorithms, available at <http://madlib.incubator.apache.org/>
- R statistical language
- SAS, in many forms, but especially with the SAS Accelerator for Greenplum
- PMML, Predictive Modeling Markup Language

## Check your knowledge

### Check your knowledge

What is a key benefit of implementing ELT (extract, load, and transform) process in the data preparation phase of the data analytics lifecycle?

OK      BACK TO LESSON HOME

DELL.COM

### Check your knowledge:

What is a key benefit of implementing ELT process in the data preparation phase of the data analytic lifecycle?

### Question:

What is a key benefit of implementing ELT process in the data preparation phase of the data analytic lifecycle?

### Answer:

---

## Check your knowledge

Check your knowledge

What is a key activity performed in the Communicate Results phase of the data analytic lifecycle?

A. Choosing an analytical technique based on the project goals      C. Sharing the analytic results with stakeholders

B. Data exploration and variable selection      D. Preparing the data repository

DELL EMC

### Check your knowledge:

What is a key activity performed in the communicate results phase of the data analytic lifecycle?

- A. Choosing an analytical technique based on the project goals
- B. Data exploration and variable selection
- C. Sharing the analytic results with stakeholders
- D. Preparing the data repository

### Question:

What is a key activity performed in the communicate results phase of the data analytic lifecycle?

### Answer:

## Check your knowledge

Check your knowledge

In which phase does the project team apply clustering, classification, or other analytic techniques?

A. Discovery      C. Model planning  
B. Model building      D. Communicate results

DELL EMC

### Check your knowledge:

In which phase does the project team apply clustering, classification, or other analytic techniques?

- A. Discovery
- B. Model building
- C. Model planning
- D. Communicate results

### Question:

In which phase does the project team apply clustering, classification, or other analytic techniques?

### Answer:

---

## Module summary

### Module summary

Key points covered in this module:

- Key phases of data analytic lifecycle
- The importance of a well-defined business problem
- Reframing a business problem as an analytics challenge
- Key deliverables of an analytics project
- The purpose of an analytics plan

# Basic data analytics methods Using R

## Introduction



**Basic data analytics methods using R**

Upon completion of this module, you should be able to:

- ✓ Write simple R code to read, process, and write datasets.
- ✓ Perform exploratory data analysis and proper data visualizations.
- ✓ Apply statistical techniques such as estimation and hypothesis testing.

DELL EMC

After completing this module, you should be able to perform basic data analytics using R, apply basic data visualization techniques using R, and use the appropriate statistical inference techniques such as estimation and hypothesis testing, also using R.

## Lesson: Introduction to the R programming language

### Introduction

# Lesson: Introduction to the R programming language

DELL EMC

### Introduction to the R programming language

This lesson covers:

- Using the RStudio graphical user interface.
- Getting data into (and out of) R.
- Data types and structures in R.



DELL EMC

After completing this lesson, you should be able to navigate through the RStudio GUI and write simple R code to import and export data out of R, as well as properly use the various R data types and structures.

## Introduction to R

### Introduction to R

- 8th ranked popular programming language (TIobe Index for January 2018)
- Open source programming language
  - Ideal for data analysis
  - Strong user community
  - Extensive package system for additional statistics and graphics functionality
- Supports data manipulation, computations, and graphical output
  - Numerous data types and structures
  - Loops and conditionals (if-then-else)
  - Input and output capabilities
  - Extensive graphics and charting capabilities
- Ideal for interactive analysis

  
Dell EMC

R is an integrated suite of software facilities for data manipulation, calculation, and visualization. R is ideal for interactive analysis, but R code may be scripted when the analysis steps must be repeated.

Reference:

[www.tiobe.com/tiobe-index/](http://www.tiobe.com/tiobe-index/)

# Basics of R programming

## Basics of R programming

- Variables do not need to be explicitly declared in advance
  - Assignment operator: "<- " or "="
  - Comments preceded by "#"
  - Most operations are function calls
  - Generic functions
    - For example, `summary()` and `plot()`
    - Behavior depends on what data type/structure is passed

```

2 # load in the sales transaction data
3 sales <- read.csv("c:/purchases.csv",
4                   header=TRUE,sep=",")
5
6 # examine the sales data
7 summary(sales)
8 cor <- cor(sales$age, sales$income)
9 plot(sales$age, sales$income,
10      main="Correlation = ", cor)
11
12 # build a logistic regression model
13 # To predict likelihood of purchase
14 model <- glm(purchase ~ income + Age,
15               data=sales,
16               family=binomial(link="logit"))
17
18 # examine model and diagnostic plots
19 summary(model)
20 plot(model)

```

第11章 项目管理

DELL.04.C

```

2 # Read in the sales transaction data
3 sales <- read.csv("c:/purchases.csv",
4                     header=TRUE,sep=",")
5
6 # Examine the sales data
7 summary(sales)
8 cor <- cor(sales$Age, sales$Income)
9 plot(sales$Age, sales$Income,|
10       main=c("correlation = ", cor))
11
12 # Build a logistic regression model
13 # to predict likelihood of purchase
14 model <- glm(Purchase ~ Income + Age,
15               data=sales,
16               family=binomial(link="logit"))
17
18 # Examine model and diagnostic plots
19 summary(model)
20 plot(model)

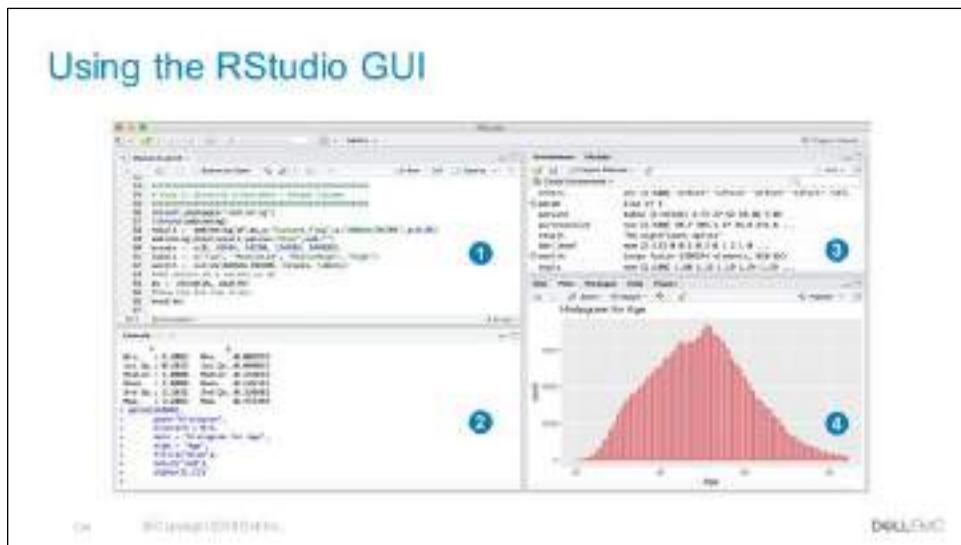
```

The sample code is provided as an example of the look and feel of the R code. Currently, the specific syntax and techniques are not important, but note how values are assigned to variables and the numerous function calls. Variables can be created when values are initially assigned. However, it may be useful to create variables in advance to explicitly define the data type and structure.

R uses several generic functions which perform different operations

depending on the input provided. For example, `summary(sales)` provides summary statistics—min, max, mean, and so on—for each column in `sales`, but `summary(model)` provides the details about the fitted logistic regression model including the estimated coefficients, their statistical significance, and several diagnostics.

## Using the RStudio GUI



The Windows version of R supports multiple GUIs. RStudio provides both a desktop and a web browser interface. This GUI is the one that you will be using in this course.

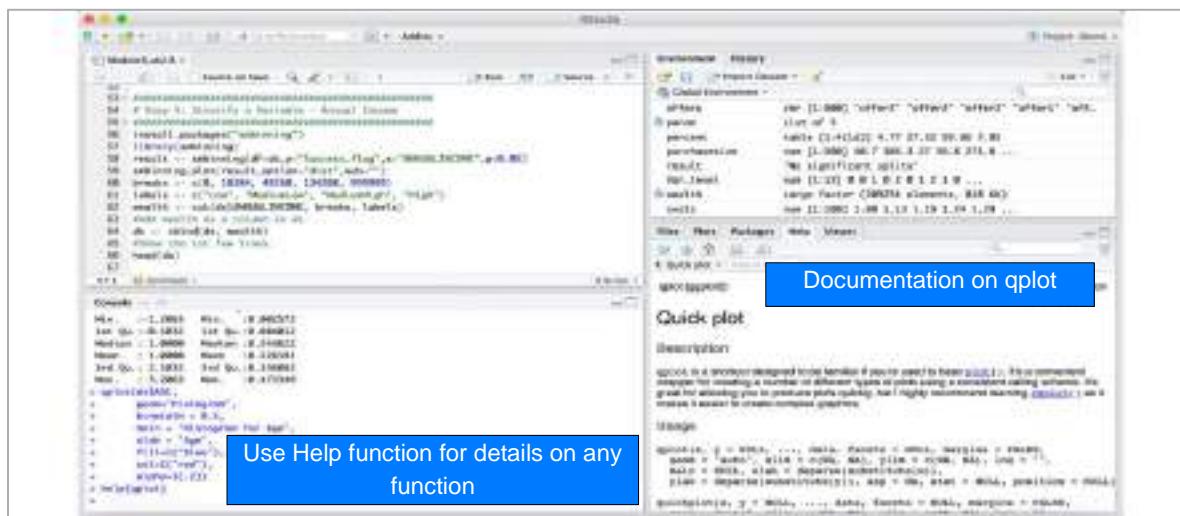
RStudio offers the following four panes:

- The upper left pane is for script editing and testing of code.
- The lower left pane is the R console, where R commands can be executed and the results displayed.
- The upper right pane provides table-oriented view of the variables stored in the current R workspace.
- The lower right provides several tabs for viewing generated plots and other details about the R environment and the Help tab.

## Using the Help command in R



The help functionality is very useful to ensure that the proper syntax is used and to examine the various options to include in the R functions. In the console, entering the **help(<topic>)** command or simply **?qplot** for this example will provide the help details on the topic.



## Importing data files into R

**Importing data files into R**

| Import Function Defaults |        |           |               |
|--------------------------|--------|-----------|---------------|
| Functions                | Header | Separator | Decimal Point |
| read.table()             | FALSE  | " "       | "."           |
| read.csv()               | TRUE   | ","       | "."           |
| read.csv2()              | TRUE   | ";"       | "."           |
| read.delim()             | TRUE   | "\t"      | "."           |
| read.delim2()            | TRUE   | "\t"      | ","           |

```

213 # importing data with read.csv
214 sales <- read.csv("c:/purchases.csv",
215                         header=TRUE,sep=",")
216
217 # importing data with read.delim
218 sales <- read.delim("c:/purchases.csv",
219                         header=TRUE,sep="\t")
220 # examine 3rd and 4th columns
221 summary(sales[,3:4])

```

```

      income          age
min. :17.00  Min. :18.00
1st Qu.:29.00  1st Qu.:32.00
Median :35.00  Median :32.90
mean   :32.49  Mean   :35.98
3rd Qu.:53.00  3rd Qu.:42.00
Max.  :98.00   Max.  :66.00

```

100 300 400 500 600 700 800 900 DELL.DIC

There are multiple functions for importing data files into R. The choice is usually based on the defaults for header, separator, and decimal point. For example, read.csv() expects the values, by default, to be comma-separated, and read.delim() expects the values to be tab delimited. Note: R uses a forward slash "/" as the separator character in pathnames for files. Thus, a file in a directory in Windows would be written as "C:/users/janedoe/My Documents/data.csv".

| Import Function Defaults |        |           |               |
|--------------------------|--------|-----------|---------------|
| Functions                | Header | Separator | Decimal Point |
| read.table()             | FALSE  | " "       | "."           |
| read.csv()               | TRUE   | " , "     | " . "         |
| read.csv2()              | TRUE   | " ; "     | " , "         |
| read.delim()             | TRUE   | " \t "    | " . "         |
| read.delim2()            | TRUE   | " \t "    | " , "         |

## Importing database tables into R

### Importing database tables into R

- R can access SQL databases
  - Establish connections
  - Process queries on the tables
- Import entire tables or a subset of records
- For large tables, any joins and complex processing are usually performed within the database

```
# Add RODBC package
install.packages("RODBC")
library(RODBC)

# Establish ODBC connection
conn <- odbcConnect("mydb",
                     uid="user",
                     pwd="password")

# Import selected records
# From sql table
hotel <- sqlQuery(conn,
                    "select
                     reserv_no,
                     hotel,
                     checkin_dt,
                     checkout_dt,
                     price
                    from
                     reservations
                    where
                     price > 150")

close(conn) #close the connection
```

© Copyright 2018 Dell Inc.

Dell EMC

In addition to importing flat data files, R can also import data directly from SQL databases. When you use R packages, such as RODBC, you can establish a connection to SQL databases and write queries to retrieve all or a subset of the data.

```
# Add RODBC package
install.packages("RODBC")
library(RODBC)

# Establish odbc connection
conn <- odbcConnect("mydb",
                     uid="user",
                     pwd="password")

# Import selected records
# From sql table
hotel <- sqlQuery(conn,
                    "select
                     reserv_no,
                     hotel,
                     checkin_dt,
                     checkout_dt,
                     price
                    from
                     reservations
                    where
                     price > 150")

close(conn) #close the connection
```

## Data types in R

### Data types in R

- Supported data types include:
  - Integers
  - Real numbers
  - Boolean or logical values (TRUE or FALSE)
  - Character
- R does not require explicit data typing of variables.
  - Good news: simplifies programming
  - Bad news: unexpected consequences may occur

#### Discussion point:

Suppose cellphone\_color is coded as follows:  
0 = black, 1 = blue, 2 = green, 3 = red, and so on

Although stored as an integer, should the data be treated as a numeric value?

140 80 minutes 100% DELL INC.

DELL INC.

As is true with most programming languages, R supports the typical data types such as integers, real numbers, Boolean values, and character data. R does not require the data type to be explicitly stated before assigning a data to a variable. Thus, writing code is greatly simplified, but caution should be taken when creating variables or reading in an external dataset into R. Regardless of how the data is represented, it is important that the data is treated properly in the analysis.

## Caution when creating variables in R



### Discussion

#### Question / Discussion Topic:

Often, numeric values can be used to denote the color of an object, such as a cellphone; but, does it make sense that red, denoted by 3, is three times more than blue, denoted by a 1?

#### Discussion Notes:

## Attribute considerations in analytics

| Attribute considerations in analytics |  |   |   |   |
|---------------------------------------|--|---|---|---|
|                                       | Categorical (Qualitative)                                      |   | Numeric (Quantitative)  |   |
|                                       | Nominal  | Ordinal   | Interval  | Ratio   |
| <b>Definition</b>                     | The values represent labels that distinguish one from another. | Attributes imply a sequence.  | The difference between two values is meaningful.                | Both the difference and the ratio of two values are meaningful. |
| <b>Examples</b>                       | ZIP codes, gender, employee IDs, TRUE or FALSE                 | Quality of diamonds, academic letter grades, magnitude of earthquakes | Temperature in Celsius or Fahrenheit, calendar dates, latitudes | Temperature in Kelvin, age, length, weight                      |
| <b>Operations</b>                     | $\neq$   | $\neq, <, >$<br>$\leq, \geq, =$                                       | $\neq, <, >$<br>$\leq, \geq, =$<br>$+$                          | $\neq, <, >$<br>$\leq, \geq, =$<br>$+$<br>$\times$              |

CC BY-SA 4.0 Dell EMC

DELL EMC

In analytics, it is useful to think of data as fitting into one of four categories: Nominal, Ordinal, Interval, or Ratio (NOIR). The choice of cellphone colors is nominal data, although the colors may be denoted by numeric values. Examples of ordinal data include {low, medium, and high} to denote, for example, the perceived risk of some event occurring. While nominal and ordinal data are both forms of categorical data that provide labels to an object's attributes, there is a clear "order" in the ordinal labels. The difference in the 2 values, however, cannot be necessarily quantified.

|                   | Categorical—Qualitative  |                              | Numeric—Quantitative                           |   |
|-------------------|--|------------------------------|--|---|
|                   | <u>Nominal</u>   | <u>Ordinal</u>               | <u>Interval</u>                                | <u>Ratio</u>  |
| <b>Definition</b> | The values represent labels that distinguish one from another. | Attributes imply a sequence. | The difference between 2 values is meaningful. | Both the difference and the ratio of 2 values are meaningful. |

## Lesson: Introduction to the R programming language

|                   |  |   |   |  |
|-------------------|--|---|---|--|
| <b>Examples</b>   | ZIP codes, gender, employee IDs, TRUE or FALSE | Quality of diamonds, academic letter grades, magnitude of earthquakes | Temperature in Celsius or Fahrenheit, calendar dates, latitudes | Temperature in Kelvin, age, length, weight |
| <b>Operations</b> | =, ≠   | =, ≠,<br><, ≤, >, ≥   | =, ≠,<br><, ≤, >, ≥,<br>+, -                                    | =, ≠,<br><, ≤, >, ≥,<br>+, -,<br>×, ÷      |

Although 20°C is greater than 10°C, it is not fair to say that 20°C is twice as hot as 10°C. Thus, these types of measurements are referred to as interval data; the differences between the values have meaning, but the ratios of the numbers may not have any true meanings. On the other hand, using the Kelvin temperature scale and the concept of absolute zero, such temperature data could be treated as ratio data. Interval and ratio types are examples of quantitative values.

Another ratio example would be a person's age. It would be fair to say that a 40-year-old person is twice as old as a 20-year-old person. Sometimes, quantitative data can be treated as qualitative data by grouping the data differently. For example, a person's age could be mapped into the following categories: {infant, child, teenager, adult, senior}. In this case, ratio data has been transformed into ordinal data.

## Common Data Structures in R

### Common Data Structures in R

- Vectors—one dimension
  - Atomic vectors
  - Lists
- Arrays—n dimensions
- Matrices—two dimensions
- Data frames
  - Similar structure to a matrix, but more like a SQL table
  - Enables access to data of various types (integer, real, character, logical)

46

DATA SCIENCE

DELL INC.

R provides several data structures to represent datasets for analysis and processing. In data analytics, vectors and data frames are commonly used, but other structures such as arrays and matrices are often used.

See the R language definition for more specifics on these topics: [cran.r-project.org/doc/manuals/r-release/R-lang.html](https://cran.r-project.org/doc/manuals/r-release/R-lang.html)

## Vectors—atomic vectors and lists

**Vectors—atomic vectors and lists**

- Atomic vectors
  - Usually just referred to as vectors
  - Contains an indexed sequence of values of the same data type such as:
    - Logical
    - Numeric
    - Character
- Lists
  - Special type of vector
  - Contains an indexed sequence of objects of different types such as:
    - Logical, numeric, and character
    - Vectors, arrays, and data frames

```

3 # example of atomic vectors
4 i <- 3
5 values <- c(2,3,4)
6 food <- c("milk", "apples", "cereal")
7
8 is.atomic(i)                      # returns TRUE
9 is.vector(i)                       # returns TRUE
10
11 # accessing members of a vector
12 food[1]                            # returns "milk"
13
14
15
16
17 # example of a list
18 purchase <- list(1,"Thomas", 534.56, TRUE, food)
19
20 is.atomic(purchase)                # returns FALSE
21 is.vector(purchase)                # returns TRUE
22 is.list(purchase)                 # returns TRUE
23
24 # accessing members of a list
25 purchase[3]                        # returns 534.56
26 purchase[[5]][[3]]                  # returns "cereal"
27

```

141      © Copyright 2018 Dell Inc.      DELL.COM

An **atomic vector** is a one-dimensional array with a single data type, often character or numeric. A special type of vector, known as a **list**, allows different types of objects to be stored in an indexed sequence. Since a list is a unique structure, based on its contents, when the literature talks about a vector, the reference is usually about an atomic vector.

Even what seems to be a scalar, such as `i <- 1`, is a vector.

The elements of a vector or a list can be accessed by providing the proper index; the indexes start at 1. For example, `food[1]` would return "milk", the first element in the vector "food". Since lists can contain fairly complex objects, the indexing is slightly more complex. For `purchase [[5]][[3]]`, the third element of the fifth element of the list, "purchase", would be returned. In "purchase", the fifth element is the vector "food", whose third element is "cereal".

```

3 # example of atomic vectors
4 i <- 1
5 values <- c(2,3,4)
6 food <- c("milk", "apples", "cereal")
7
8 is.atomic(i)                      # returns TRUE
9 is.vector(i)                       # returns TRUE
10
11 # accessing members of a vector
12 food[1]                            # returns "milk"
13
14
15
16
17 # example of a list
18 purchase <- list(1,"Thomas", 534.56, TRUE, food)
19
20 is.atomic(purchase)                # returns FALSE
21 is.vector(purchase)                # returns TRUE
22 is.list(purchase)                 # returns TRUE
23
24 # accessing members of a list
25 purchase[3]                        # returns 534.56
26 purchase[[5]][[3]]                  # returns "cereal"
27

```

## Arrays

### Arrays

- N-dimensional
- Indexed sequence of values of the same data type such as:
  - Logical
  - Numeric
  - Character

```
14 # build a 2 dimensional array
15 # with rows=3, cols=4, and pages=2
16 revenue <- array(0, dim=c(3,4,2))
17
18 # examine structure of array
19 str(revenue) # returns num [1:3, 1:4, 1:2] 0...
20
21 # assign a values to the array
22 revenue[1,1,2] <- 5
23 revenue[2,1] <- c(6,7,8)
24
25 revenue
```

, , 1

|       |      |      |      |      |
|-------|------|------|------|------|
|       | [,1] | [,2] | [,3] | [,4] |
| [1,1] | 0    | 6    | 0    | 0    |
| [2,1] | 0    | 7    | 0    | 0    |
| [3,1] | 0    | 8    | 0    | 0    |

, , 2

|       |      |      |      |      |
|-------|------|------|------|------|
|       | [,1] | [,2] | [,3] | [,4] |
| [1,1] | 5    | 0    | 0    | 0    |
| [2,1] | 0    | 0    | 0    | 0    |
| [3,1] | 0    | 0    | 0    | 0    |

Multidimensional arrays can be created with array(). In the example, a three-row by four-column by two-page cube structure is created with initial values of 0. Such a structure may be useful for representing revenue data across three geographic regions, over four quarters, for the last two fiscal years.

```
14 # build a 3 dimensional array
15 # with rows=3, cols=4, and pages=2
16 revenue <- array(0, dim=c(3,4,2))
17
18 # examine structure of array
19 str(revenue) # returns num [1:3, 1:4, 1:2] 0...
20
21 # assign a values to the array
22 revenue[1,1,2] <- 5
23 revenue[2,1] <- c(6,7,8)
24
25 revenue
```

, , 1

|       |      |      |      |      |
|-------|------|------|------|------|
|       | [,1] | [,2] | [,3] | [,4] |
| [1,1] | 0    | 6    | 0    | 0    |
| [2,1] | 0    | 7    | 0    | 0    |
| [3,1] | 0    | 8    | 0    | 0    |

, , 2

|       |      |      |      |      |
|-------|------|------|------|------|
|       | [,1] | [,2] | [,3] | [,4] |
| [1,1] | 5    | 0    | 0    | 0    |
| [2,1] | 0    | 0    | 0    | 0    |
| [3,1] | 0    | 0    | 0    | 0    |

## Matrices

**Matrices**

- 2-dimensional array
- For numeric matrices, R provides common matrix operations:
  - Transpose – `t()`
  - Multiplication – `%*%`
  - Determinant – `det()`

```

52 # build a 26 row x 2 column matrix
53 # 1st column - position in the alphabet
54 # 2nd column - letter of the alphabet
55 letter_mat <- matrix(c(1:26,letters),
56                         nrow=26, ncol=2)
57
58 # display the first three rows of the matrix
59 # what is the data type of the first column?
60 head(letter_mat,3)

```

|      | [,1] | [,2] |
|------|------|------|
| [1,] | "1"  | "a"  |
| [2,] | "2"  | "b"  |
| [3,] | "3"  | "c"  |

DELL INC.

Two-dimensional arrays are known as **matrices**. In the example, a 26-row and two-column matrix is created where the second column contains the lower-cased letter of the alphabet and the first column contains the position of the letters in the alphabet. Since matrices, as well as vectors and arrays, must contain the same data type, numeric data and character data cannot both exist in the created matrix. Thus, the numeric data is **coerced** into character data.

```

52 # build a 26 row x 2 column matrix
53 # 1st column - position in the alphabet
54 # 2nd column - letter of the alphabet
55 letter_mat <- matrix(c(1:26,letters),
56                         nrow=26, ncol=2)
57
58 # display the first three rows of the matrix
59 # what is the data type of the first column?
60 head(letter_mat,3)

```

|      | [,1] | [,2] |
|------|------|------|
| [1,] | "1"  | "a"  |
| [2,] | "2"  | "b"  |
| [3,] | "3"  | "c"  |

## Data frames

### Data frames

- 2-dimensional data structure
- Columns are easily referenced by name
- Different columns may have different data types

```
67 # Build data frame based on two vectors
68 # 1st. column - position in the alphabet
69 # 2nd column - letter of the alphabet
70 seq <- c(1:26)
71 letter_df <- data.frame(seq,
72                         letters,
73                         stringsAsFactors=FALSE)
74
75 class(letter_df)          # returns "data.frame"
76
77 # display the first three rows of the data.frame
78 # what is the data type of the first column?
79 #                                     of the second column?
80 head(letter_df,3)
```

```
seq letters
1 1 a
2 2 b
3 3 c
```

"data.frame": 26 obs. of 2 variables:  
\$ seq : int 1 2 3 4 5 6 7 8 9 10 ...  
\$ letters: chr "a" "b" "c" "d" ...

Dell EMC

When different data types must be stored in a single R variable, a **data frame** is often useful. A data frame has a matrix-like structure but is very similar to a SQL table structure. In a data frame, the columns can be of different data types, such as character or numeric. Also, it is common to create column names, such as seq and letters, in the example.

Especially for data frames with many columns, the use of column names eliminates the need for explicitly stating that the 43rd column, for example, should be accessed. Thus, letter\_df\$letters would reference the stored letters of the alphabet and can be treated as a vector. Using str() to examine the structure of letter\_df, reveals that the object is a data frame and the stored data types of integer and character.

```
67 # Build data frame based on two vectors
68 # 1st. column - position in the alphabet
69 # 2nd column - letter of the alphabet
70 seq <- c(1:26)
71 letter_df <- data.frame(seq,
72                         letters,
73                         stringsAsFactors=FALSE)
74
75 class(letter_df)          # returns "data.frame"
76
77 # display the first three rows of the data.frame
78 # what is the data type of the first column?
79 #                                     of the second column?
80 head(letter_df,3)
```

```
seq letters
1 1 a
2 2 b
3 3 c
```

"data.frame": 26 obs. of 2 variables:  
\$ seq : int 1 2 3 4 5 6 7 8 9 10 ...  
\$ letters: chr "a" "b" "c" "d" ...

## Factors

### Factors

- A factor is a variable with specific levels
  - For example, low, medium, and high discount
- Useful for analyzing categorical data
- Two kinds of factors
  - Unordered for nominal data
  - Ordered for ordinal data
- By default, `data.frame()` treats character data as factors

```

66: # build data frame based on two vectors
67: # 1st column - position in the alphabet
68: # 2nd column - letter of the alphabet
69: seq <- c(1:26)
70: letter_df <- data.frame(seq,
71:                           letters)
72:
73: # examine structure of data frame
74: str(letter_df)

'data.frame': 26 obs. of 2 variables:
$ seq   : int 1 2 3 4 5 6 7 8 9 10 ...
$ letters: Factor w/ 26 levels "a","b","c","d",...

```

```

76: # explicitly structure as ordinal data
77: letter_df$letters <- ordered(letter_df$letters)
78:
79: # examine modified structure
80: str(letter_df)

'data.frame': 26 obs. of 2 variables:
$ seq   : int 1 2 3 4 5 6 7 8 9 10 ...
$ letters: Ord.factor w/ 26 levels "a" $\leq$ "b" $\leq$ "c" $\leq$ "d" $\leq$ ...

```

In statistical experiments, factors are those attributes to be examined at various levels. For example, if the effect of temperature at different levels on the yield of a manufacturing process is examined in an experiment, the factor could possibly be the three levels of 190°C, 200°C, and 210°C. Since most character data is treated as categorical data, `data.frame()` treats character data, by default, as factors. These factors can be unordered or ordered, depending on if the data is nominal or ordinal, respectively.

```

66: # build data frame based on two vectors
67: # 1st column - position in the alphabet
68: # 2nd column - letter of the alphabet
69: seq <- c(1:26)
70: letter_df <- data.frame(seq,
71:                           letters)
72:
73: # examine structure of data frame
74: str(letter_df)

'data.frame': 26 obs. of 2 variables:
$ seq   : int 1 2 3 4 5 6 7 8 9 10 ...
$ letters: Factor w/ 26 levels "a","b","c","d",...

```

```

76: # explicitly structure as ordinal data
77: letter_df$letters <- ordered(letter_df$letters)
78:
79: # examine modified structure
80: str(letter_df)

'data.frame': 26 obs. of 2 variables:
$ seq   : int 1 2 3 4 5 6 7 8 9 10 ...
$ letters: Ord.factor w/ 26 levels "a" $\leq$ "b" $\leq$ "c" $\leq$ "d" $\leq$ ...

```

## Exporting files and graphics out of R

### Exporting files and graphics out of R

- Comparable functions to the import functions
  - `write.table()`
  - `write.csv()`
  - `write.csv2()`
  - `write.delim()`
  - `write.delim2()`
- Graphics can be exported

```
153: # create a new jpeg file for plot
154: jpeg(file="c:/data/hist.jpg")
155:
156: # create the histogram ||
157: hist(salesage)
158:
159: # close off the graphic device
160: dev.off()
```

DELL DSC

There are comparable write functions to the various read functions. Also, plots, charts, and other graphics can be exported to files. In this example, a .jpg file is created. In R, this .jpg file becomes the current output device. When the histogram is created, the graphics output is directed to the .jpg file. After you close the current graphic device, the output is directed back to the Plot tab in RStudio.

```
153: # create a new jpeg file for plot
154: jpeg(file="c:/data/hist.jpg")
155:
156: # create the histogram ||
157: hist(salesage)
158:
159: # close off the graphic device
160: dev.off()
```

DELL DSC

## Check your knowledge

### Check your knowledge

- In data analytics, what does the acronym NOIR represent?
- Why is NOIR important in data analytics?
- What is a benefit of a data frame over a matrix?



44 © 2018 Dell Inc.

DELL.COM

## Check your knowledge



### Discussion

## Question / Discussion Topic:

1. In data analytics, what does the acronym NOIR represent?
2. Why is NOIR important in data analytics?
3. What is a benefit of a data frame over a matrix?

## Discussion Notes:

## Lesson summary

### Lesson summary

This lesson covered the following topics:

- Using the RStudio Graphical User Interface
- Getting data into and out of R
- Data types and structures in R



10

© Copyright 2018 Dell Inc.

Dell EMC

## Lesson: Analyzing and exploring data

### Introduction

# Lesson: Analyzing and exploring data

DELL EMC

### Analyzing and exploring data

This lesson covers:

- The importance of visualization.
- Examining a single variable.
- Examining pairs of variables.
- Indications of dirty data.



DELL EMC

This lesson addresses the importance of proper data visualizations in the data analytics.

## What is data visualization?

What is data visualization?

**United States Census Bureau**

The presentation of statistics with images that depict the meaning of the statistics. – census.gov

**Bloomberg Businessweek**

Data visualization schematically abstracts information to bring about a deeper understanding of the data, wrapping it in an element of awe. – Bloomberg Business Week

**Information Visualization: Design for Interaction**

The use of computer-supported, interactive, visual representations of abstract data to amplify cognition. – Card et al., 1999

10      © Copyright 2018 Dell Inc.      DELL.COM



Here are three different definitions of data visualization:

The US Census Bureau defines data visualization as a term that means “the presentation of statistics with images that depict the meaning of the statistics.”<sup>1</sup>

Bloomberg Business Week says “Data visualization schematically abstracts information to bring about a deeper understanding of the data, wrapping it in an element of awe.”<sup>2</sup>

A better definition is perhaps from Card et al.’s book in 1999. They define data visualization as “the use of computer-supported, interactive, visual representations of abstract data to amplify cognition.”<sup>3</sup>

**References:**

<sup>1</sup>Yau, N. Visualizing Census Data. *The United States Census Bureau.*  
[census.gov/library/video/data\\_visualization1.html](https://www.census.gov/library/video/data_visualization1.html)

<sup>2</sup>Popova, M. (2012). Data Visualization: Stories for the Information Age.  
[www.bloomberg.com/news/articles/2009-08-12/data-visualization-stories-for-the-information-age](https://www.bloomberg.com/news/articles/2009-08-12/data-visualization-stories-for-the-information-age)

<sup>3</sup>Card, S. K., Mackinlay, J. D., & Shneiderman, B. (Eds.). (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

## Why is data visualization important?

### Why is data visualization important?

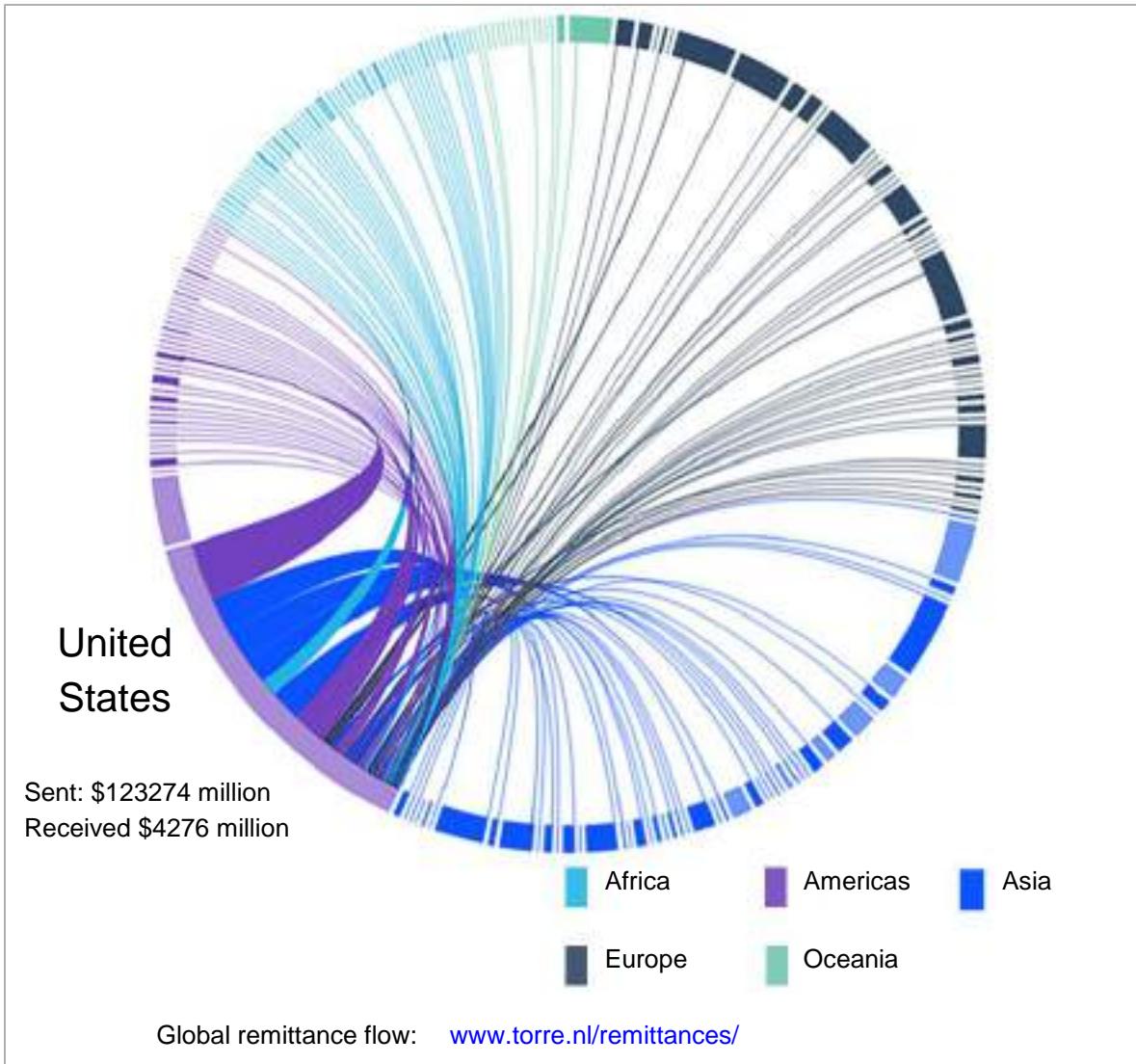
- We are visual beings
  - Sight is a key sense for information understanding
  - We have been using visuals for many centuries
- Data is easier to read in visual form
- Helps discover new knowledge
- Applies to any domain
- Assist in analysis and communication

Data is easier to read in visual form

Source: Dell EMC

“One look is worth a thousand words,” as the Chinese proverb puts it. Humans have been using visuals for many, many centuries. In prehistoric times, ancient humans already knew how to use cave drawing to communicate. The oldest known cave art comes from the Cave of El Castillo in Cantabria, Northern Spain, dated back to at least 40,800 years.

Data is often much, much easier to read in a visual form and it can also help discover new knowledge. For example, a histogram of household income data makes it much easier for us to identify the distribution of the income, compared to reading a gigantic Excel spreadsheet.



Data visualization helps with analysis and communication. The image shows the global movements of money due to remittances—money migrants send to their home countries—with the United States migrants highlighted. You can view the visualization at: [www.torre.nl/remittances/](http://www.torre.nl/remittances/). The graph shows that most money from the US went to Asia (yellow), Americas (green), and Europe (dark blue).

## Lesson: Analyzing and exploring data

# Anscombe's Quartet

## Anscombe's Quartet

| Property   | Values |
|--|--------|
| Mean(x) = 9  |        |
| Excellence of fit: r=0.00                                  |        |
| Excellence of fit: r=0.98                                  |        |
| Excellence of fit: r=0.99                                  |        |
| Correlation coefficient: r=0.00                            |        |
| Linear regression line: Y=3.00+2.00x, r=0.9999999999999999 |        |

| x    | y    | x    | y    | x    | y    | x    | y     |
|------|------|------|------|------|------|------|-------|
| 1.00 | 7.08 | 3.00 | 9.14 | 1.00 | 8.04 | 9.00 | 12.90 |
| 3.00 | 7.08 | 3.00 | 9.14 | 3.00 | 8.04 | 9.00 | 12.90 |
| 3.00 | 9.14 | 3.00 | 9.14 | 3.00 | 8.04 | 9.00 | 12.90 |
| 3.00 | 9.14 | 3.00 | 9.14 | 3.00 | 8.04 | 9.00 | 12.90 |
| 1.00 | 7.08 | 1.00 | 7.08 | 1.00 | 7.08 | 1.00 | 7.08  |
| 3.00 | 7.08 | 3.00 | 9.14 | 3.00 | 8.04 | 9.00 | 12.90 |
| 3.00 | 9.14 | 3.00 | 9.14 | 3.00 | 8.04 | 9.00 | 12.90 |
| 3.00 | 9.14 | 3.00 | 9.14 | 3.00 | 8.04 | 9.00 | 12.90 |
| 1.00 | 7.08 | 1.00 | 7.08 | 1.00 | 7.08 | 1.00 | 7.08  |
| 3.00 | 7.08 | 3.00 | 9.14 | 3.00 | 8.04 | 9.00 | 12.90 |
| 3.00 | 9.14 | 3.00 | 9.14 | 3.00 | 8.04 | 9.00 | 12.90 |
| 3.00 | 9.14 | 3.00 | 9.14 | 3.00 | 8.04 | 9.00 | 12.90 |

Anscombe's Quartet is a synthesized example created by the statistician F. J. Anscombe.

| Property                                  | Values  |
|---|---|
| Mean of x in each case                    | 9   |
| Exact variance of x in each case          | 11  |
| Exact mean of y in each case              | 7.5 (to 2 d.p)                                    |
| Variance of Y in each case                | 4.13 (to 2 d.p)                                   |
| Correlations between x and y in each case | 0.816   |
| Linear regression line in each case       | $Y = 3.00 + 0.500x$<br>(to 2 d.p and 3 d.p resp.) |

## Are these datasets identical?



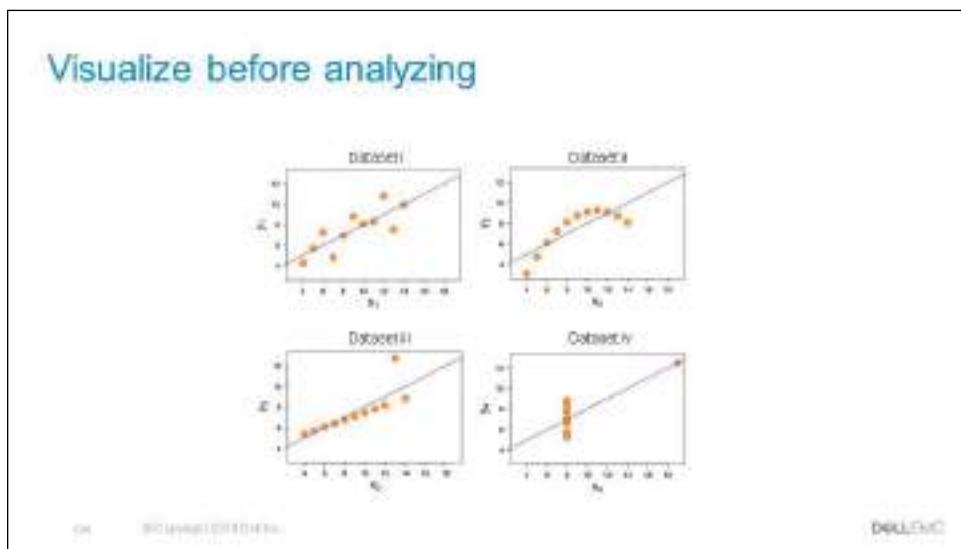
## Discussion

**Question / Discussion Topic:**

Look at the properties and values of these four datasets. Based on standard statistical measures of mean, variance, and correlation—our descriptive statistics—these datasets are identical. Or, are they?

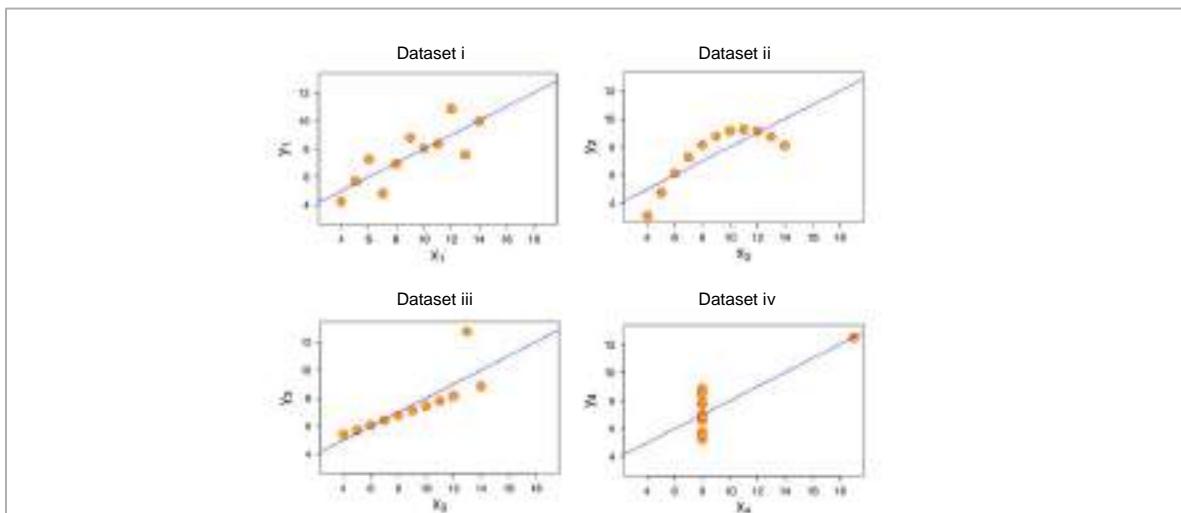
**Discussion Notes:**

## Visualize before analyzing



If you visualize each data set using a scatterplot and a regression line superimposed over each plot, the datasets seem quite different:

- Dataset i is the best candidate for a regression line, although there is much variation.
- Dataset ii is definitely nonlinear.
- Dataset iii is a close match, but over predicts at higher value of x and has an extreme outlier.
- For dataset iv, there is some question on whether the computed regression line is a good representation of the data for any possible value x.



## Examining distribution of a single variable

### Examining distribution of a single variable

Multiple ways to visualize a single variable:

- Plot (variable)
- Hist (variable)
- Plot(density(variable))
- Rug Plot - provides distribution of variable along x and y axis



DATA SOURCE: DELL INC.

Here is an example of distribution of maturity balance—mortgage balance that is up for renewal is maturity balance. As shown in the graph, most of the mortgages are from about zero to somewhat less than a million.

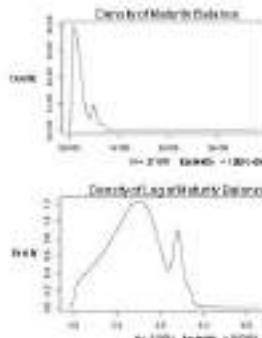
The rug() function creates a one-dimensional density plot, as well: notice how it emphasizes the area under the curve. Notice that, at the bottom of the graph, the rug provides the distribution of the number of balances across the scale.

An example of the jitter function is also shown in the graph in light blue. It is the same as rug, but with a few more features that could be explored using the help function.

## Density plots—what to look for

### Density plots—what to look for

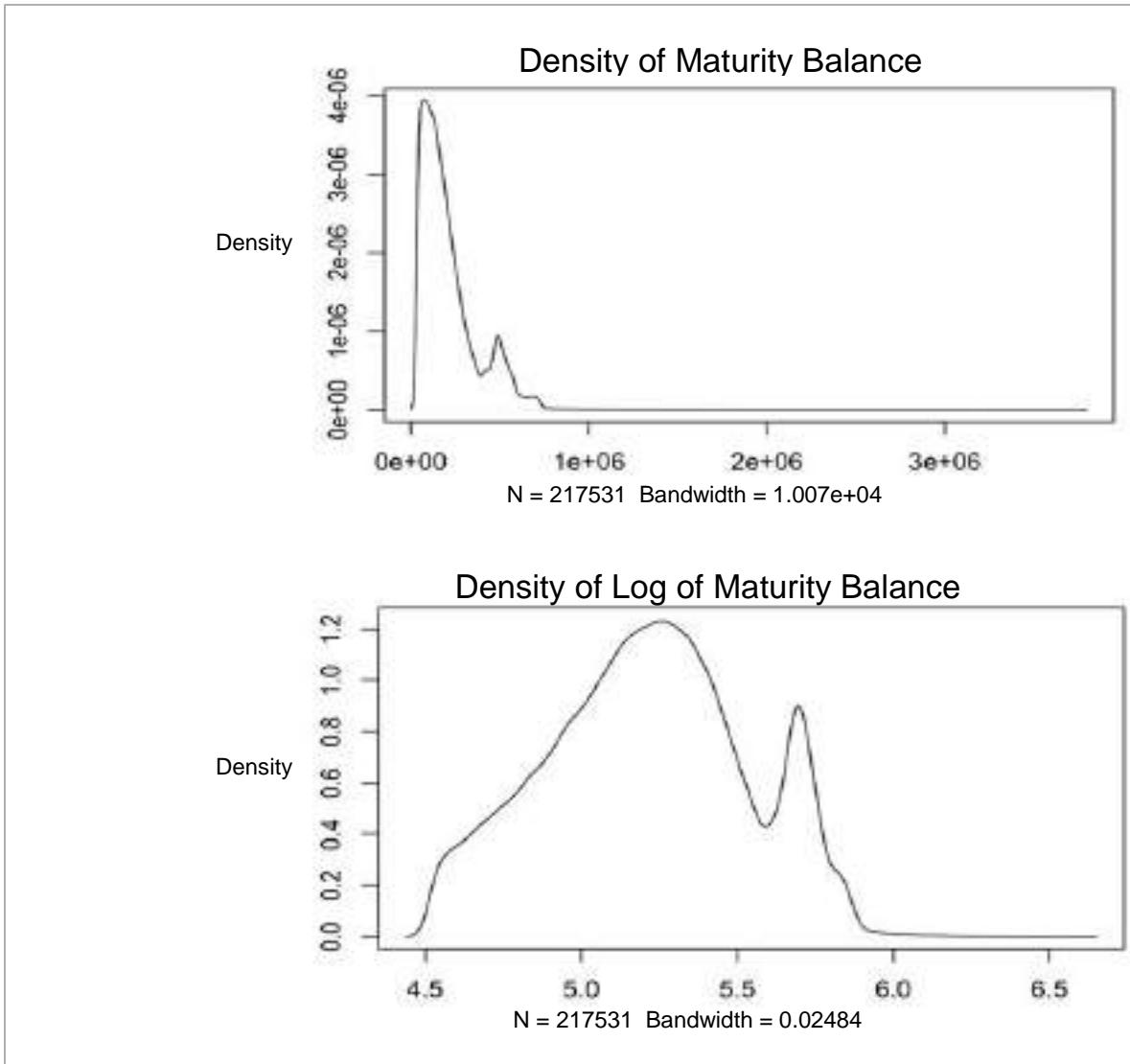
- Shape of the distribution
  - Unimodal? Bimodal?
  - Any long tails?
  - Approximately normal?
- Outliers or anomalies
  - Possibly evidence of dirty data
- Example – density of maturity balance
  - Range from 0 to 3 million
  - Plotting log of data gives better sense of distribution



100 300 1000 3000 10000 30000

Dell EMC

During the data exploration phase, it is important to understand the spread of the data and whether the values are strongly concentrated in a certain range. If the data is skewed, viewing the log of the data—if it is all positive—can help you detect structure that you might otherwise miss in a regularly scaled graph.



See if the data is unimodal or multimodal—that gives you an idea of how many distinct populations, with distinct behavior patterns, might be mixed into your overall population. Knowing if the data is approximately normal, or can be transformed to approximately normal—for example, by taking the log—is important, since many modeling techniques assume that the data is approximately normal in distribution. Look for obvious signs of dirty data—outliers or unlikely-looking values.

For this example, look at the density plot of maturity balances, in \$ CAD, of customers at a Financial institution. The range here is extremely wide, from around close to \$0 CAD to over \$3,000,000 CAD. Extreme ranges such as this one are typical of monetary data, including income, customer value, tax liabilities, bank account sizes, and so on. In fact, all of this kind of data is often assumed to be distributed lognormally—that is, its log is a normal distribution.

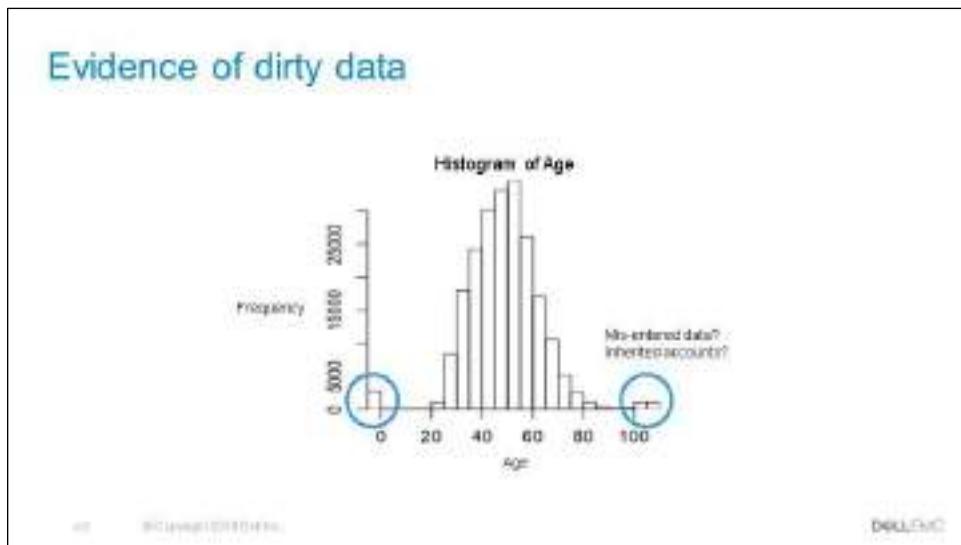
## Lesson: Analyzing and exploring data

The data range makes it hard for you to see much detail, so take the log of it, and then density plot it. In this example, as seen in the graph, there is a segment that has higher maturity balances as compared to the normal distribution, resulting in two spikes that are magnified with the use of density on Log of maturity balance.

CODE:

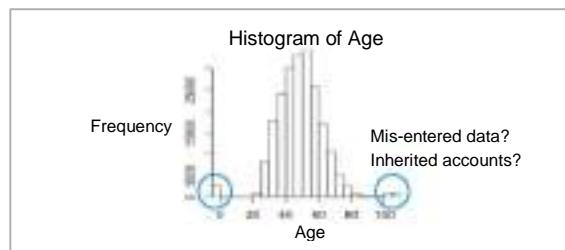
```
plot(density(ds$maturingbalance), main= "Density of Maturity Balance")
plot(density(log10(ds$maturingbalance)), Main= "Density of Log of Maturity
Balance")
```

## Evidence of dirty data



Here is an example of how dirty data might manifest itself in your visualizations. You are looking at the age distribution of account holders at the bank. Mean age is about 50, approximately normally distributed.

You see a few accounts with an age less than or equal to 0, which is not possible. This issue could be caused by missing data or misentered data that must be corrected.



The customers who are older than 100 are possibly also misentered data. Or, these examples could be accounts that have been passed down to the heirs of the original account holders, and not updated. You may want to exclude these account holders or at least threshold the age to be considered in the analysis.

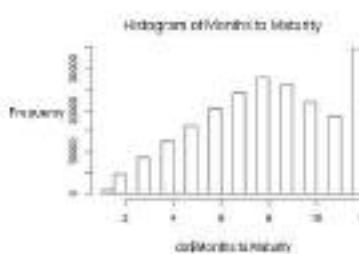
CODE:

```
hist(ds$missingAge, main= "Histogram of Age (adjusted)")
```

## Saturated data

### Saturated data

- Do we really have no mortgages with maturity more than 12 months after observation date?
- Or was there an error in entry which constrained the maturity date to 12 months from now?

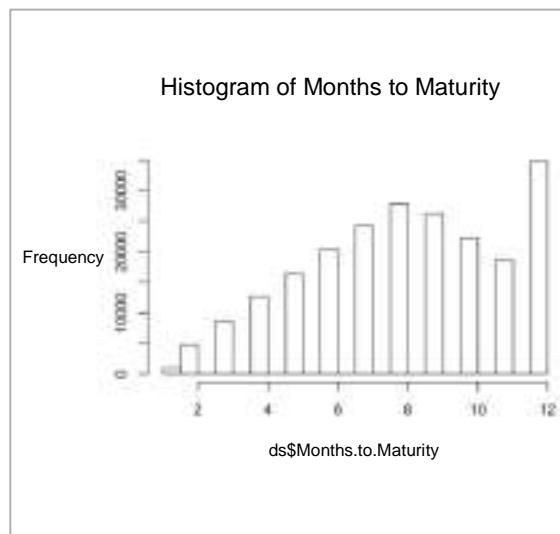


A histogram titled "Histogram of Months to Maturity". The x-axis is labeled "ds\$Months.to.Maturity" and ranges from 0 to 12 with major ticks every 2 units. The y-axis is labeled "Frequency" and ranges from 0 to 30,000 with major ticks every 10,000 units. The distribution is skewed right, with the highest frequency occurring at 12 months (approximately 35,000). Other frequencies are lower, peaking around 8 months.

100 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800 3000

DELL EMC

Here is another example of dirty—or, at least, incompletely documented—data. You are looking at the months to maturity for mortgages. This is a column provided in the dataset.



The first thing you notice is that you do not seem to have mortgages with maturity more than 12 months from observation. You also notice that you have a disproportionate number of loans maturing around 12 months, which is not in line with the distribution of loans in the other months.

What would you do about this variance? If you are analyzing probability of default,

## Lesson: Analyzing and exploring data

it is probably safe to eliminate the data—or keep the assumption that maturity will not be after 12 months. You must go back to the source data and check if there was a constraint applied when you are retrieving the source data.

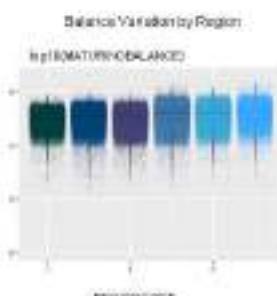
CODE:

```
hist(ds$Months.to.maturity, main= "Histogram of Months to Maturity")
```

## Analyzing relationship between two variables

### Analyzing relationship between two variables

- Two continuous variables (or two discrete variables)
  - Scatterplots
  - Linear models: graph the correlation
  - Stripplots, hexbin plots
    - = More legible: color-based plots for high-volume data
- Continuous vs. discrete variable
  - Dot, box-and-whisker plots, dotplot or barchart
- Example:
  - Mortgage balance initiation by region code
  - Scatterplot with jitter, with box-and-whisker overlaid
  - Maturity balance equally distributed across all regions



Balance Variation by Region

LOG(MATURITYBALANCE)

REGIONCODE

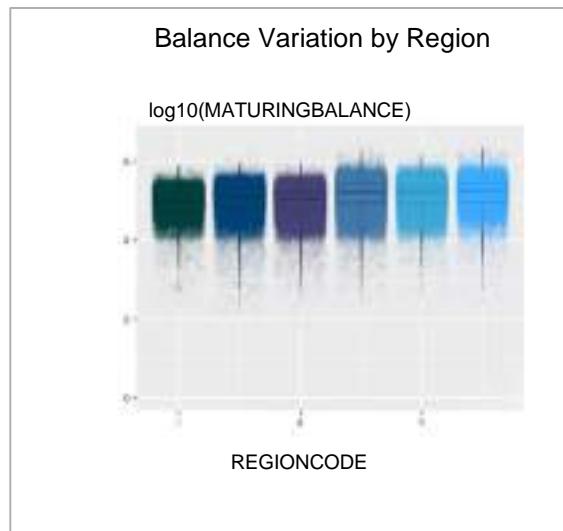
DOLLAR/C

As you noticed with Anscombe's quartet, scatterplots are a good first visualization for the relationship between two variables, especially two continuous variables.

For very high-volume data, scatterplots are problematic; with too much data on the page, the details can get lost. Sometime the jitter() function can create enough uniform variation to see the associations more clearly.

There are other alternatives for plotting continuous vs. discrete variables. Dotplots and barcharts plot the continuous value as a function of the discrete value when the relationship is one-to-one. Box and whisker plots show the distribution of the continuous variable for each value of the discrete variable.

The example here is of logged maturity balances as a function of region—first digit of the zip. Logged, in this case, means data that uses the logarithm of the value instead of the value itself. In this example, you have also plotted the scatterplot beneath the box-and-whisker, with some jittering so each line of points widens into a strip.



The "box" of the box and whisker shows the range that contains the central 50 percent of the data; the line inside the box is the location of the median. The "whiskers" give you an idea of the entire range of the data. Usually, box and whiskers also show "outliers" that lie beyond the whiskers, but they are turned off in this graph. This graph shows how maturity balances do not vary by region. The graph is in ggplot, which is fairly complicated. The commands are:

```
library(ggplot2)
# the outlier.size=0 prevents the boxplot from plotting the outlier
ggplot(data, aes(x=RegionCode, y=log10(MaturityBalance))) +
  geom_boxplot(outlier.size=0, alpha=0.1) + points
# plot the jittered scatterplot, color-code the points
  geom_point(aes(colour=RegionCode), alpha=0.02, position="jitter").
```

You can read more about ggplot2 at [had.co.nz/ggplot2/](http://had.co.nz/ggplot2/)

## Two variables—what to look for

Two variables—what to look for

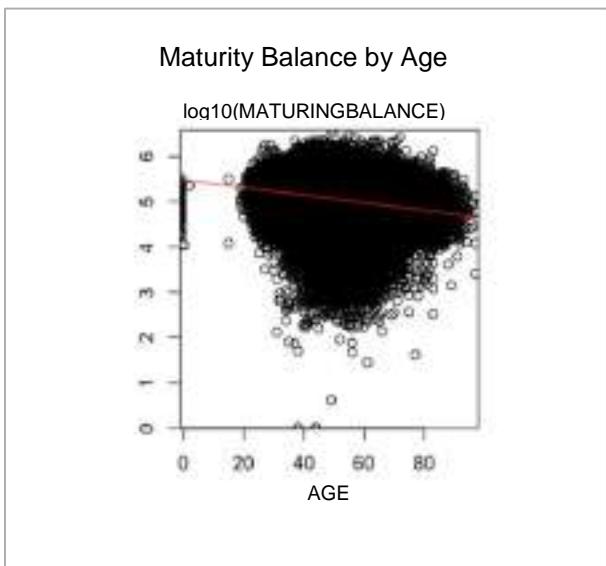
Is there a relationship between the two variables?

- Linear? Quadratic?
- Exponential?
  - Try semi-log or log-log plots
- Concentrated? Multiple clusters
- Example
  - Scatterplot
  - Red line: linear fit

In the example here, the relationship seems approximately linear; we have plotted the regression line in red. Sometimes, a standard regression line just does not capture the relationship. In this case, the `loess()` function in R—also `lowess()`—fits a nonlinear line to the data. Here, the `loess` curve is drawn in blue.

R-Code

Assume a dataset named `ds` with variables `MaturityBalance` and `Age`. The R code to generate the plot shown here is as follows:

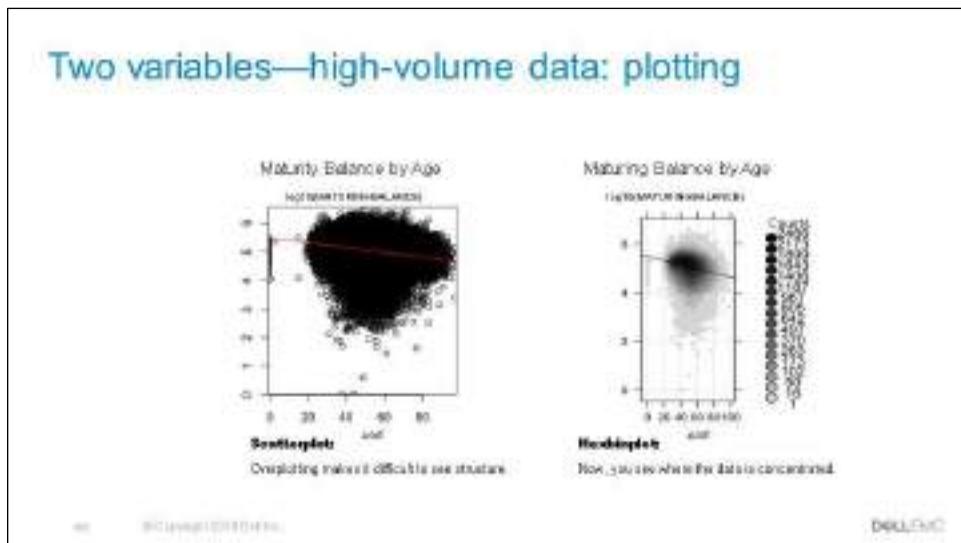


```
with(ds,
{
  plot(log10(MaturityBalance) ~ Age)

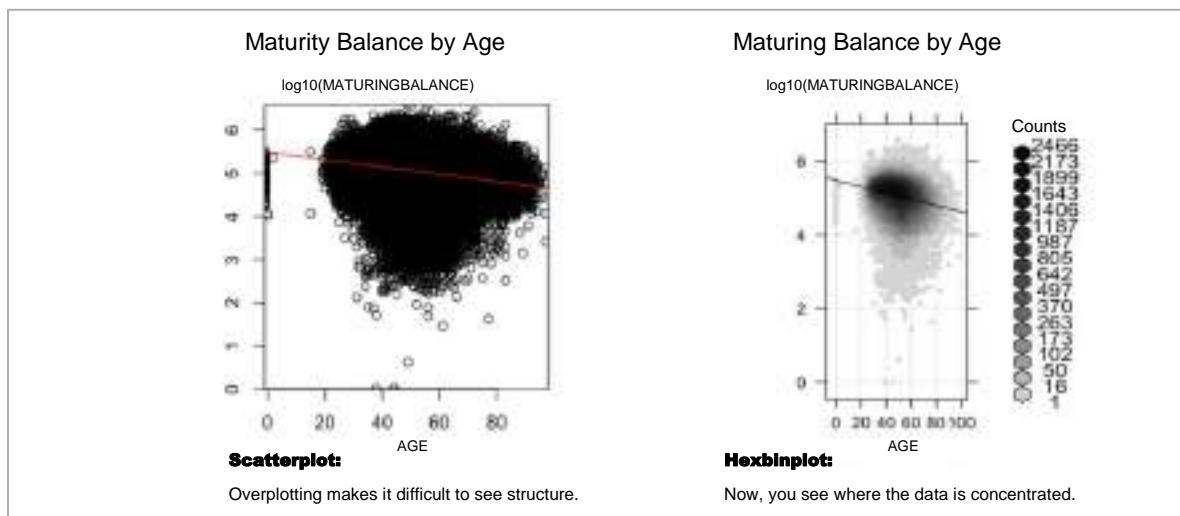
abline(lm(plot(log10(MaturityBalance)
~ Age), lcol="red"))

lines(lowess(plot(log10(MaturityBalanc
ce) ~ Age), lcol="blue"))
})
```

## Two variables—high-volume data: plotting



When you have too much data, the structure becomes difficult to see in a scatterplot. Here, you are plotting logged Maturity Balance against Age. The "blob" that you get on the scatterplot on the left suggests a linear relationship. However, you cannot really see where the data points are concentrated.



On the right, the same data is plotted using a hexbinplot. Hexbinplots are similar to two-dimensional histograms, where shading tells us how populated the bin is. Now, you can see that the data is more densely clustered in a streak that runs through the center of the plotted points, roughly along the regression line. The biggest concentration is around 30 to 50 years of age.

## Lesson: Analyzing and exploring data

```
library(hexbin)

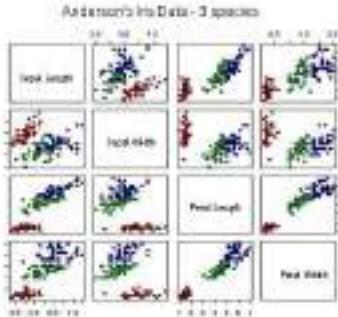
# "g" adds the grid, "r" the regression line
# sqrt transform on the count gives more dynamic range to the shading

hexbinplot(log10(MaturityBalance) ~ Age,
           data=zcta, trans = sqrt, inv = function(x) x^2,
           type=c("g", "r"))
```

## Establishing multiple pairwise relationships between variables

**Establishing multiple pairwise relationships between variables**

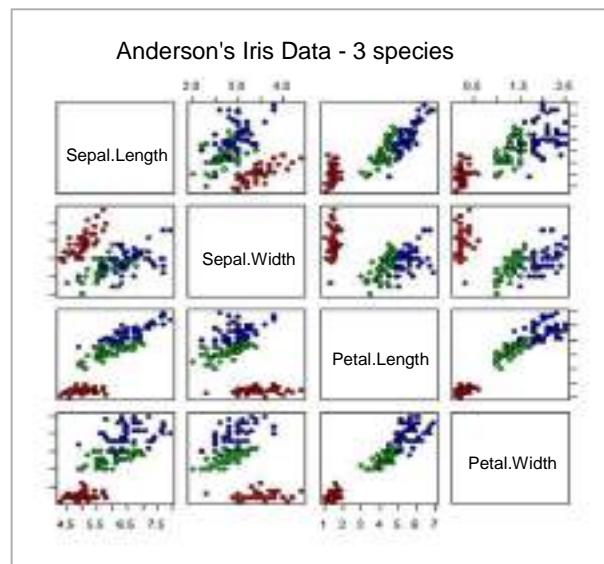
- Why?
  - Examine many two-way relationships quickly
- How?
  - `pairs()` can generate a plot of each pair of variables
- Example
  - Iris characteristics
    - Strong linear relationship between petal length and width
    - Petal dimensions discriminate species more strongly than sepal dimensions



The figure shows a 4x4 grid of scatter plots for Anderson's Iris Data. The diagonal elements represent individual variable distributions (Sepal.Length, Sepal.Width, Petal.Length, Petal.Width). The off-diagonal elements represent pairwise relationships between variables. The data points are colored by species: red for Setosa, green for Versicolor, and blue for Virginica. The plots illustrate that petal dimensions (Length and Width) are more effective for species discrimination than sepal dimensions.

At times, it is useful to see multiple values of a dataset in context, to visually represent data relationships so as to magnify differences or to show patterns hidden within the data that summary statistics do not reveal. In the graphic represented here, the variables sepal length, sepal width, petal length, and petal width are compared with three species of irises. The key is not listed in the graphic. Colors are used to represent the different species, allowing you to compare differences across species for a particular combination of variables.

Consider the values encoded in the second square from the upper right, where sepal length is compared with petal length. Values for petal length are encoded across the bottom; values for sepal length are encoded on the right side of the graphic. This image shows that the green and blue species are well matched, although the blue species has longer petals in the main. The petal length for the red species, however, remains markedly the same, and varies only in the lower half of sepal length values.



## Relationship between sepal length and sepal width



### Discussion

#### Question / Discussion Topic:

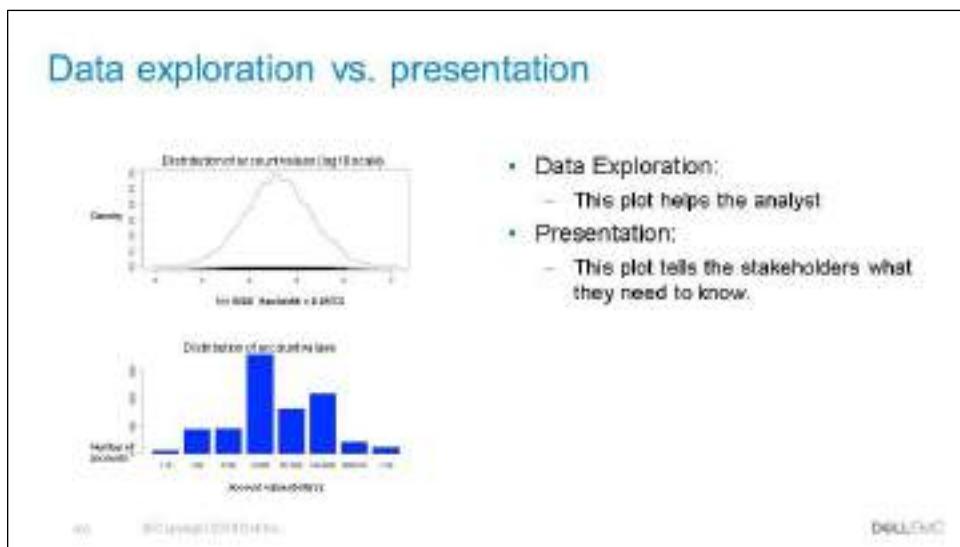
As an exercise, imagine fitting a regression line to each of these individual graphs. What would you make of the relationship between sepal length and sepal width?

Using the iris dataset included with base R, the R code for generating the plot is:

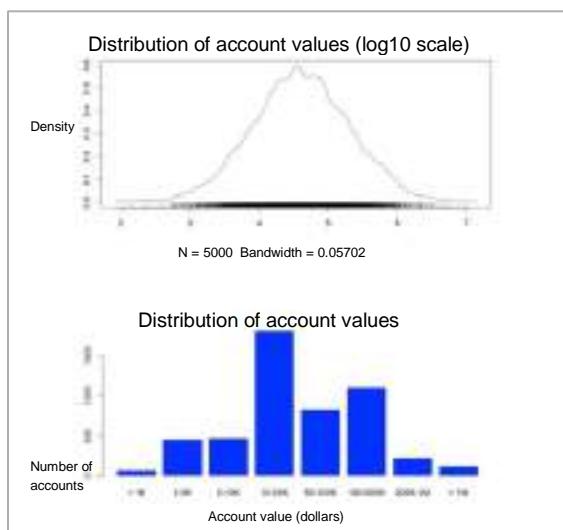
```
pairs(iris[1:4], main = "Anderson's Iris Data -- 3 species",
      pch = 21, bg = c("red", "green3", "blue")[unclass(iris$Species)] )
```

#### Discussion Notes:

## Data exploration vs. presentation



Finally, let us touch on the difference between using visualization for data exploration, and for presenting results to stakeholders. The presented plots and tips are designed to make the details of the data as clear as possible for the data scientist to see structure and relationships. These technical graphs do not always effectively convey the information that must be conveyed to nontechnical stakeholders. For them, you want crisp graphics that focus on the message you want to convey.



It would be hard to explain the density plot to stakeholders. For one thing, density plots are fairly technical, and for another, it is awkward to explain why you are logging the data before showing it. You can convey essentially the same information by partitioning the data into "log-like" bins, and presenting the histogram of those bins, as shown in the bottom plot. Here, you can see that the bulk of the accounts are in the 1000 to 1M range, with the peak concentration in the 10 K to 50 K range, extending out

to about 500 K. This representation gives the stakeholders a better sense of the customer base than the top graphic would.

## Lesson: Analyzing and exploring data

**Note:** The reason that the lower graph is not symmetric like the upper graph is because the bins are only "log-like". They are not truly log10 scaled. Log10-scaled bins would be closer to 1 K to 3 K, 3 K to 10 K, 10 K to 30 K, and so on.

Plot for the top graphic:

```
plot(density(log10(income), adjust=0.5), main="Distribution of account values  
(log10 scale)")  
rug(log10(income))
```

Plot for the bottom graphic:

```
# create "log-like bins"  
breaks = c(0, 1000, 5000, 10000, 50000, 100000, 5e5, 1e6, 2e7)  
# bin and label the data  
bins = cut(income, breaks, include.lowest=T,  
          labels = c("< 1K", "1-5K", "5-10K", "10-50K", "50-100K", "100-500K", "500K-1M",  
          "> 1M"))  
# plot the bins.  
plot(bins, main = "Distribution of account values", xlab = "account value (dollars)",  
      ylab = "number of accounts", col="blue")  
e chose, however, might seem more "natural" to the stakeholders.
```

## Check your knowledge

### Check your knowledge

1. In the Iris slide example, how would you characterize the relationship between sepal width and sepal length?
2. Did you notice the use of color in the Iris slide? Was it effective? Why or why not?



Dell EMC

## Check your knowledge



### Discussion

## Question / Discussion Topic:

1. In the Iris slide example, how would you characterize the relationship between sepal width and sepal length?
2. Did you notice the use of color in the Iris slide? Was it effective? Why or why not?

## Discussion Notes:

## Lesson summary

### Lesson summary

This lesson covered the following topics:

- The importance of visualization
- Examining a single variable
- Examining pairs of variables
- Indications of dirty data



40

DATA SCIENCE

DELL INC.

## Lesson: Statistics for model building and evaluation

### Introduction

# Lesson: Statistics for model building and evaluation

DELL EMC

### Statistics for model building and evaluation

This lesson covers:

- Estimation
- Hypothesis testing
- Significance and power
- Statistics in the data analytics lifecycle

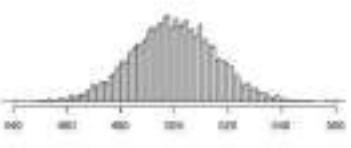


DELL EMC

## Statistical inference—drawing conclusions based on data

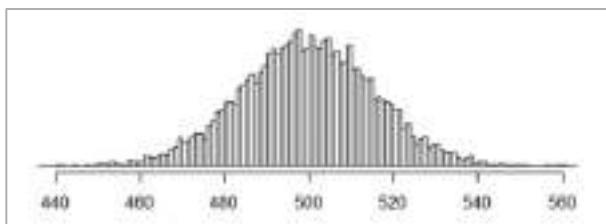
Statistical inference—drawing conclusions based on data

- Estimation
  - Estimate a population characteristic, such as:
    - o Mean
    - o Variance
    - o Percentiles
  - Typical approaches:
    - o Point estimation
    - o Confidence intervals
- Hypothesis testing
  - Evaluate an assertion about populations of interest
    - o Example: Is the mean of Population A different than the mean of Population B?
  - Common techniques:
    - o t-test (assumes normal distribution)
    - o Wilcoxon Rank-Sum (non-parametric)
    - o Analysis of Variance (ANOVA)



DATA/DOC

Statistical inference is the process of drawing conclusions based on data. Two forms of statistical inference are estimation and hypothesis testing.



In estimation, the value of some population characteristic must be determined. For example, what is the mean? What is the variance? What percentage of customers must wait in line at the bank more than 10

minutes before seeing a teller? Often, determining the point estimates of the mean or the variance is fairly straightforward or formulaic. However, as may be intuitively obvious, the accuracy of any point estimate depends on the sample size. One would expect a sample size of 100 to provide better estimates than a sample size of 5. It depends on what is being estimated. Confidence intervals are often used to communicate the uncertainty in the point estimates.

Hypothesis testing is often used to assess if there has been a fundamental shift. For example, such assertions may be:

- A new exercise program is more effective than the existing program at reducing a patient's weight.
- Customers offered incentive A will purchase more than customers offered incentive B.

## Normal distribution

**Normal distribution**

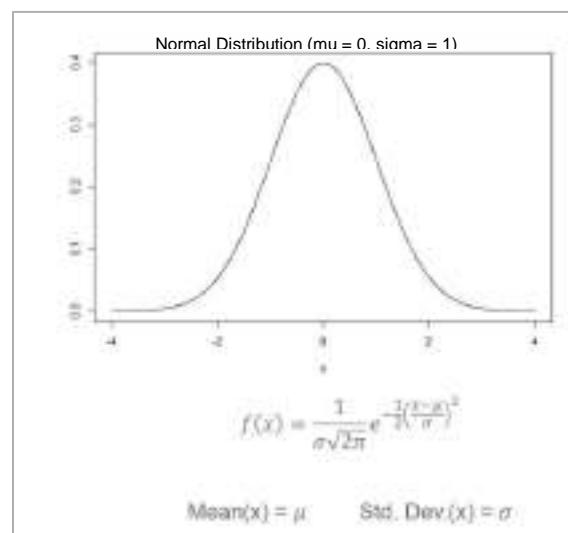
- Useful to describe many datasets
- Assumed in many statistical and modeling techniques
  - Population of interest
  - Random error terms (noise)
- Discussion: What are good estimates for  $\mu$  and  $\sigma$ ?

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Mean( $x$ ) =  $\mu$       Std. Dev( $x$ ) =  $\sigma$

CC BY-SA 3.0 Dell Inc.

The normal distribution, also known as the Gaussian distribution or colloquially as the “bell curve”, applies to many datasets. In many statistical techniques, it is common to assume that the sample data was from a normally distributed population.



In some modeling approaches, such as linear regression, the random error terms or noise is assumed to be normally distributed. The mean and standard deviation of normally distributed population are the parameters  $\mu$  and  $\sigma$ , respectively. If you want to better understand a normal distributed population, these parameters must be estimated based on a random sample.

## Point estimation—normal distribution parameters, $\mu$ and $\sigma$

Point estimation—normal distribution parameters,  $\mu$  and  $\sigma$

For  $\mu$ , use the sample mean:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

where  $n$  is the sample size

For  $\sigma$ , use the sample standard deviation:  $s$

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Discussion

What other point estimates may provide reasonable estimates?  
How accurate are these point estimates?

100% [Data Science and Big Data Analytics v2](#) [Dell EMC](#)

For a random sample of size  $n$ , from a normal distribution,  $\mu$  is estimated by the arithmetic mean of the sample, also referred to as the sample mean.  $\sigma$  is estimated by the square root of the sample variance, which is known as the sample standard deviation. Because of certain statistical properties, these estimates are often the preferred estimates but are not the only possible point estimates. Regardless of which estimates may be used, there will always be some uncertainty in the estimates, based on sample data.

### Point estimation



#### Discussion

### Question / Discussion Topic:

1. What other point estimates may provide reasonable estimates?
2. How accurate are these point estimates?

**Discussion Notes:**

## Confidence intervals

### Confidence intervals

Used to convey the uncertainty in the point estimates.

A  $100(1-\alpha)\%$  confidence interval for the mean of a normal distribution

$$\bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

where  $t_{\alpha/2, n-1}$  is the upper  $\alpha/2$  quantile of the t-distribution with  $n-1$  degrees of freedom.

For a 95% confidence interval, choose:  $\alpha = 0.05$

Interpretation: In repeated random sampling, 95% of the intervals will straddle the value of the true, but unknown mean (95% confidence in any interval).

```
9 # simulate 100 random normal values
10 # mu = 10 and sigma = 4
11 set.seed(524423)
12 v <- rnorm(100,10,4)
13
14 # estimate mu and sigma
15 mean(v) # returns 10.17
16 sd(v) # returns 3.91
17
18 # determine the t-value
19 delta <- qt(.025,.99,lower.tail=FALSE) *
20 sd(v)/sqrt(100)
21 delta # returns -0.34
22
23 # calculate 95% confidence interval on mu
24 c(mean(v)-delta,
25 mean(v)+delta) # returns 9.83 to 10.51
```

The general approach is to calculate an interval (LB, UB) such that, in repeated random samples, a specified percentage of the intervals straddle the true but unknown parameter of interest. Thus, if 1,000 random samples of size n were chosen from a normal population, one would expect 950—95 percent—of the 1,000 corresponding confidence intervals to straddle the true but unknown mean.

Note: LB denotes the lower bound, and UB denotes the upper bound.

For the estimate of a mean from a normal population, the uncertainty is the product of a t-distribution value and the sample standard deviation divided by the square root of the sample size.

In R, the t value can be obtained with `qt()`, the quantile function for the t-distribution, and the corresponding degrees of freedom,  $n - 1$ , and the desired  $100(1 - \alpha)\%$  confidence level.

## Motivation for hypothesis testing

### Motivation for hypothesis testing

- Manufacturing example
  - A manufacturing process has been producing parts with a mean diameter of 10 mm and a standard deviation of 0.2 mm.
  - It is suspected that the manufacturing process mean has shifted.
- The approach
  - Assume the process is producing parts with a mean diameter of 10 mm.
  - Sample the parts and measure their diameters.
  - If the average diameter of these parts is significantly different than 10 mm:
    - o Then conclude the process has shifted.
    - o Or else conclude the process has not shifted.
- Challenges
  - How is "significantly" determined?
  - What is the risk of erroneously concluding that the process has shifted, when it has not?
- Discussion: Would confidence intervals help?

CC BY-SA 4.0 - Open Data Institute

DELL INC.

## Hypothesis testing



### Discussion

## Question / Discussion Topic:

Would confidence intervals help?

## Discussion Notes:

## The t-test on the mean ( $\mu=10$ )

**The t-test on the mean ( $\mu=10$ )**

```

Null hypothesis ( $H_0$ ):  $\mu=10$ 
Alternative hypothesis ( $H_A$ ):  $\mu \neq 10$ 
* Inputs
  - Sample data vector
  - Null hypothesis
  - Confidence level for conf.interval
* P-value – significance of the test
  - Small p-values (say <0.05) support  $H_A$ 
  - Larger p-values support  $H_0$ 
* Interpretation of p = 0.3209
  - If  $\mu=10$ , then the observed sample mean and std. dev. for the 500 obs., would be expected ~32% of the time.
  - Thus, accept (do not reject)  $H_0$ 

```

```

28 # simulate 500 random normal values
29 # mu = 10 and sigma = 4
30 set.seed(524423)
31 v <- rnorm(500,10,4)
32
33 #estimate mu and sigma
34 mean(v)      #returns 10.17
35 sd(v)        #returns 3.91
36
37 t.test(v, mu=10, conf.level=0.95)

```

```

One Sample t-test

data: v
t = 0.9936, df = 499, p-value = 0.3209
alternative hypothesis: true mean is not equal to 10
95 percent confidence interval:
 9.830345 10.516817
sample estimates:
mean of x
10.17358

```

In this example, a simulated random sample of 500 observations were generated with a mean of 10 and a standard deviation of 4. In fact, this dataset is the same one that was used in the confidence interval example. The parameter estimates were close to these values, but somewhat different.

Using the `t.test()` function, a hypothesis test is performed to evaluate whether the population mean is different than 10. The p-value provides the probability that the t-value of 0.9936, which is based on the sample mean, sample standard deviation, and sample size, would be obtained under the assumption that the true mean equals 10. Since p is close to 0.32, one would expect such results 32 percent of the time if the mean was 10. Thus, it would not be prudent to reject the null hypothesis.

Note that the provided confidence interval is equivalent to the confidence interval generated earlier.

## The t-test on the mean ( $\mu=9.7$ )

**The t-test on the mean ( $\mu=9.7$ )**

Null hypothesis ( $H_0$ ):  $\mu=9.7$   
Alternative hypothesis ( $H_A$ ):  $\mu \neq 9.7$

- \* Inputs
  - Sample data vector
  - Null hypothesis
  - Confidence level for conf. interval
- \* P-value – significance of the test
  - Small p-values (say <0.05) support  $H_A$
  - Larger p-values support  $H_0$
- \* Interpretation of p = 0.006942
  - If  $\mu=9.7$ , then the observed sample mean and std. dev. for the 500 obs., would be expected~0.7% of the time.
  - Thus, reject  $H_0$  in favor of  $H_A$

```

46: t.test(x, mu=9.7, alt="two.sided", level=0.99)
47: 
48: 
49:   t-test

50: 
51: data: x
52: t = -3.7338, df = 499, p-value = 0.006942
53: alternative hypothesis: true mean is not equal to 9.7
54: 95 percent confidence interval:
55: 9.698453-10.008817
56: sample estimates:
57: mean of x
58: 10.17798

```

DELL DMC

If the null hypothesis is  $\mu=9.7$  for the same dataset, the resulting p-value is 0.006942. Since the observed p-value is quite small, less than 7 out of 1000, either something truly rare occurred for a mean equal to 9.7, or the mean does not equal 9.7. In this case, at a 99.3 percent confidence level, the null hypothesis would be rejected in favor of the alternative hypothesis.

## Possible errors in hypothesis testing

**Possible errors in hypothesis testing**

| Decision              | H <sub>0</sub> is true | H <sub>0</sub> is false |
|-----------------------|------------------------|-------------------------|
| Accept H <sub>0</sub> | Correct outcome        | Type II error           |
| Reject H <sub>0</sub> | Type I error           | Correct outcome         |

- + Analyst controls the likelihood of committing a Type I error.
  - Set the significance level,  $\alpha$ , to a small enough value (e.g.  $\alpha=0.05$ ).
  - If H<sub>0</sub> is true, the analyst will only reject H<sub>0</sub> with probability  $\alpha$ .
- + Analyst influences the likelihood of committing a Type II error.
  - Probability,  $\beta$ , of committing a Type II error depends on the significance level,  $\alpha$ .
  - Choose a large enough sample size to detect a specified effect size.
    - For H<sub>0</sub>:  $\mu=10$ , successively larger sample sizes will be required to detect a true mean of 11, 10.5, 10.1, 10.01, or 10.001.
    - The respective effect sizes are 1, 0.5, 0.1, 0.01 or 0.001.

DRAFT DOCUMENT

In hypothesis testing, there are two types of errors that can occur. A Type I error occurs when H<sub>0</sub> is rejected but H<sub>0</sub> is actually true. A Type II error occurs when H<sub>0</sub> is accepted but H<sub>0</sub> is actually false.

| Decision              | H <sub>0</sub> is true | H <sub>0</sub> is false |
|-----------------------|------------------------|-------------------------|
| Accept H <sub>0</sub> | Correct outcome        | Type II error           |
| Reject H <sub>0</sub> | Type I error           | Correct outcome         |

The power of a test is the probability of correctly rejecting the null hypothesis. It is denoted by  $1-\beta$ , where  $\beta$  is the probability of a type II error. Because the power of a test improves as the sample size increases, power is used to determine the necessary sample size. In the difference of means, the power of a hypothesis test depends on the true difference from the null hypothesis. In other words, for a fixed significance level, a larger sample size is needed to detect a smaller difference from the assumed value of the mean. In general, the magnitude of the difference is known as the effect size. As the sample size becomes larger, it is easier to detect a given effect size.

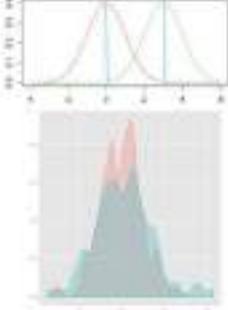
## Hypothesis tests for comparing multiple populations

**Hypothesis tests for comparing multiple populations**

**Welch's t-test:**  
 Tests if the means of two populations are equal  
 $H_0: \mu_1 = \mu_2 = 0$   
 $H_a: \mu_1 - \mu_2 \neq 0$   
 Assumes the populations are normally distributed

**Wilcoxon Rank Sum test:**  
 Tests if one population is shifted to the right or left of the other population  
 No normality assumption (nonparametric)  
 Uses the rank sum statistic of the observed sample values

**Analysis of Variance (ANOVA):**  
 Tests if the means of k populations are equal  
 $H_0: \mu_1 = \mu_2 = \dots = \mu_k$   
 $H_a: \mu_1 \neq \mu_2 \neq \dots \neq \mu_k$   
 Assumes the populations are normally distributed



DATA SOURCE: © 2018 Dell Inc.

In addition to hypothesis tests conducted on samples from one population, hypothesis tests can be conducted to compare characteristics of multiple populations. Welch's t-test examines the difference between two population means, based on samples drawn from the populations. When the normality assumption does not seem to be true, the Wilcoxon rank sum test can be used. Since no underlying distribution or parameterization is assumed, such a test is referred to as a nonparametric test. ANOVA is a useful test to compare the means from multiple populations and has extensive application in design of experiments.

## Comparing Welch's t-test and Wilcoxon rank sum

### Comparing Welch's t-test and Wilcoxon rank sum

|   |   |
|---|---|
| <p>Welch's t-statistic: <math>t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}</math></p> <p>If <math>t</math> is close to zero, do not reject <math>H_0</math>. Use the p-value.</p> <p>Each population is assumed to be normally distributed.</p> <p>Handles unequal variances.</p> <p>Allows unequal sample sizes.</p> | <p><b>Wilcoxon rank sum procedure</b></p> <ol style="list-style-type: none"><li>1. Order the <math>n_1 + n_2</math> observations.</li><li>2. Assign ranks:<ul style="list-style-type: none"><li>- 1 to the smallest value</li><li>- 2 to the 2nd smallest value, ...</li><li>- <math>n_1 + n_2</math> to the largest value<ul style="list-style-type: none"><li>o If the two populations are the same, the ranks should be somewhat uniformly assigned across the samples.</li><li>o If the two populations are shifted, the lower ranks should be somewhat more assigned to one sample than the other sample.</li></ul></li></ul></li><li>3. Sum the ranks for one sample (denoted <math>W</math>).</li><li>4. Determine the probability of observing <math>W</math> or a greater value, under <math>H_0</math> assumption of no shift.<ul style="list-style-type: none"><li>- The p-value</li></ul></li></ol> |
|---|---|

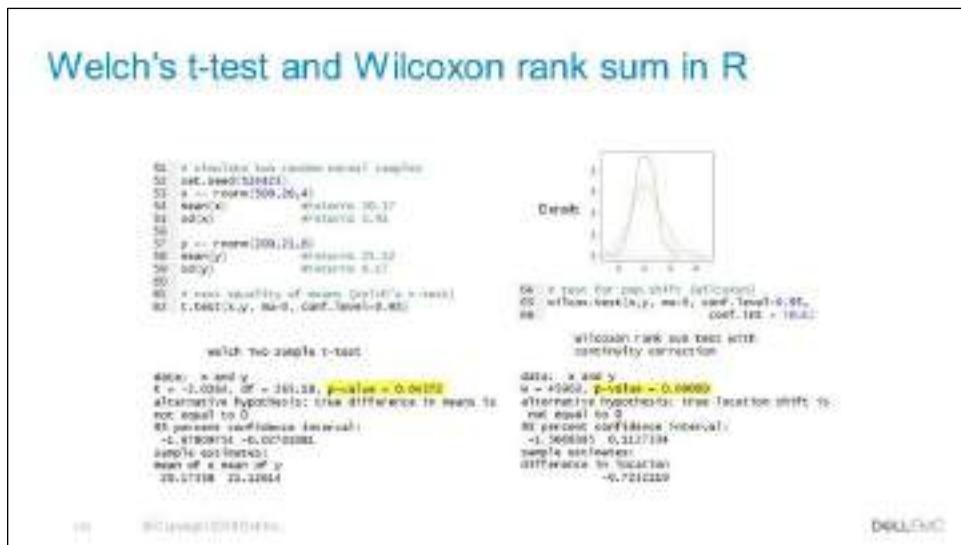
Welch's t-test is very similar to t-test for one population mean. In the Welch's t-test, the difference in the sample means is adjusted by a function of the sample variances and respective sample size. If the Welch's t-statistic is close to zero, then do not reject  $H_0$ ; again, use the p-value to determine how rare the calculated t-statistic was.

The Wilcoxon rank sum test ranks the observed values from the two samples. If the two populations are the same, the ranks should be somewhat evenly assigned to the observations. However, if the two populations are shifted, the lower ranks would be assigned to one sample and the larger ranks will be assigned to the other sample.

The Wilcoxon rank sum test then considers whether the sum of the ranks, denoted  $W$ , from one sample is unusually small or large based on the assumption that both samples were from the same population. The probability of observing the value  $W$  or something to a greater extreme is the corresponding p-value for the Wilcoxon rank sum test.

As the next example will illustrate, if the assumptions of the t-test are true, the t-test is more likely to detect a shift in the means than the Wilcoxon rank sum test.

## Welch's t-test and Wilcoxon rank sum in R



In this example, two random samples of sizes 500 and 200 are randomly generated. The first sample is taken from a normal population with a mean of 20 and a standard deviation of 4. The second sample is from a normal distribution with a mean of 21 and a standard deviation of 6. Thus, at a 95 percent confidence level—significance level of 0.05—Welch's t-test would reject H<sub>0</sub>, that the difference of the means is zero. However, the Wilcoxon rank sum test would not reject the null hypothesis for the same significance level.

For both hypothesis tests in R, similar outputs are provided. For the observed samples, the t statistics and rank sum statistic, W, are reported with their corresponding p-values as well as the respective alternative hypotheses. Again, illustrating the connection between hypothesis tests and confidence intervals, the 95% confidence interval for t-test does not straddle zero which corresponds to rejecting the null hypothesis at a 0.05 significance level.

For the Wilcoxon rank sum test, the 95% confidence interval does straddle zero, which corresponds to not rejecting the null hypothesis at a 0.05 significance level. Of course, the choice of Welch's t-test or Wilcoxon rank sum test depends on which assumptions appear to hold true.

For the t-test, it may be necessary to transform the two samples to meet the normality assumption. When such transformations are not possible, the Wilcoxon rank sum test should be considered especially when the evidence of outliers exists.

## Lesson: Statistics for model building and evaluation

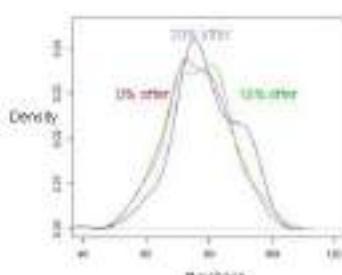
Such extreme values can significantly affect the sample means and standard deviations and the t-test.

By assigning ranks to the observed values, the Wilcoxon rank sum test is less sensitive to extreme values. For example, the smallest value is assigned a rank of 1 and the largest value is assigned a rank of  $n_1 + n_2$ .

## Analysis of variance (ANOVA)

**Analysis of variance (ANOVA)**

- Test the equality of means of  $k$  populations  
 $H_0: \mu_1 = \mu_2 = \dots = \mu_k$   
 $H_a: \mu_i \neq \mu_j$  for at least one pair of  $i, j$
- Useful in experiments where the levels of one or more factors are adjusted
- Example: Determine the effect of discount on average purchase amount
  - Factor: discount
  - Three levels of discount: 0%, 10%, and 20%
  - Randomly assign discounts to customers
- Is there a significant shift in the mean purchase amount?



Density

Purchase

0% offer

10% offer

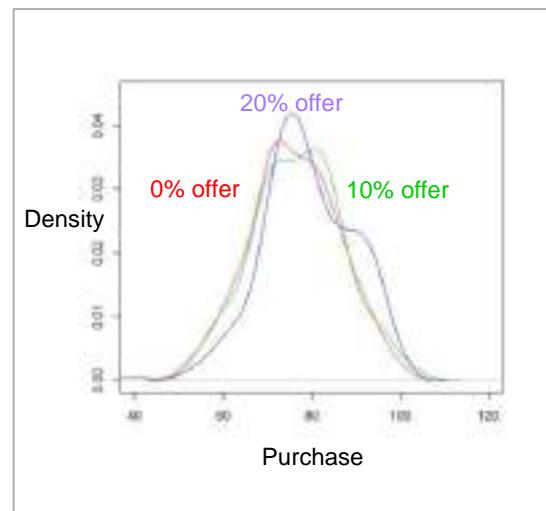
20% offer

DELL INC.

Analysis of Variance (ANOVA) is a generalization of the hypothesis test for equality of means. Here, you have multiple populations, and you want to see if any of the population means are different from the other means. That means that the null hypothesis is that ALL the population means are equal to each other. The alternative hypothesis is that at least two of the means are not equal.

An example, suppose that everyone who visits our retail website either gets one of two promotional offers, or no promotion at all. You want to see if making the promotional offers makes a difference. In this example, only one factor, promotion, is considered with three levels of discount: 0%, 10%, and 20%. So, this technique is known as one-way ANOVA.

You can do multiway ANOVA (MANOVA), as well. For instance, if you want to analyze offers and day of week simultaneously, that would be a two-way ANOVA. Multiway ANOVA is usually accomplished by doing a linear regression on the outcome, using each of the categorical treatments as an input variable.



## Analysis of variance (ANOVA) in R

### Analysis of variance (ANOVA) in R

- Using `aov()`, model the purchase amount as a function of offer
- Examine the observed variances
  - For the 3 offer means: 342.2
  - For the 500 obs. within each level: 102.7
  - Under  $H_0$ , expect variance to be equal.
- F-statistic: 3.332
  - Variation among sample means / variation within groups
  - The ratio of the two variances
  - With degrees of freedom = (k-1, 450-k)
- p-value: 0.0366
  - Reject  $H_0$  at 0.05 significance level
  - Do not reject  $H_0$  at 0.01 signif. level

```

01 # display dataset offer data
02 # collected in a data frame
03 summary(offer_dt)
04
05 offer      purchase
06 offer9:129   min. :39.03
07 offer10:150  1st Qu.:70.20
08 offer11:162  median :76.89
09          mean  :78.99
10          3rd Qu.:83.91
11          max. :103.04
12
13 # perform analysis of variance
14 results <- aov(purchase ~ offer,
15   data=offer_dt)
16 summary(results)
17
18 Df Sum Sq Mean Sq F value Pr(>F)
19 offer       2    684    342.2  3.331 0.0366 *
20 Residuals 447 45831   102.7
21
22 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

```

Using the `aov()` function in R, a hypothesis test on the equality of several means can be performed. Under the assumption of equality of means, the mean square errors for the offers and the observations within the groupings would be equal to the population variance. Thus, the resulting ratio of these values would be close to 1. Furthermore, this ratio is known to follow an F-distribution. So, the likelihood of such an F-statistic can be determined and is provided as the p-value of 0.0366.

## Tukey honest significant differences

### Tukey honest significant differences

- ANOVA identifies a difference in means
  - But which means are significantly different?
- Tukey HSD provides confidence intervals
- Significant difference between offers 0% and 20%

```
## calculate confidence intervals
tukeyHSD(results, conf.level=0.95)
```

```
Tukey multiple comparisons of means
95% Family-wise confidence level

Fit: aovChisq ~ purchase ~ offer, data = offer_0f

$offer
            diff      lwr      upr      p adj
offer10-offer0  0.8188920 -0.4632416 1.512953  0.6837719
offer20-offer0  2.3923634  0.2827260 3.120212  0.0234180
offer30-offer0  2.3430830 -0.8146170 3.033344  0.3833951
```

If ANOVA indicates that not all of the means are equal, the Tukey honest significant differences can be used to generate a family of confidence intervals at the specified confidence level.

## Statistics in data analytics lifecycle



### Discussion

#### Question / Discussion Topic:

### Statistics in data analytics lifecycle



- Model planning and building phases
  - Can I predict the outcome with the inputs that I have?
  - Which inputs can be used?
  - Is the model accurate?
  - Does the model perform better than "the obvious guess"?
  - Does the model perform better than another candidate model?
- Operationalize
  - Does the model make a difference?
    - o Are we preventing customer churn?
    - o Have we raised profits?
  - What are areas for improvement?

As data scientists, you use statistical techniques not only within your modeling algorithms but also during the early model building stages, when you evaluate your final models and when you assess how your models improve the situation when deployed in the field.

#### Discussion Notes:

## Check your knowledge

### Check your knowledge

1. An estimate of the population mean is obtained; how can the uncertainty in that estimate be expressed?
2. If the normality assumption for a hypothesis test does not appear to be true, what are possible options?



DELL EMC

## Check your knowledge



### Discussion

#### Question / Discussion Topic:

1. An estimate of the population mean is obtained; how can the uncertainty in that estimate be expressed?
2. If the normality assumption for a hypothesis test does not seem to be true, what are possible options?

#### Discussion Notes:

## Lesson summary

### Lesson summary

This lesson covered the following topics:

- Estimation
- Hypothesis testing
- Significance and power
- Statistics in the data analytics lifecycle



100

© Copyright 2018 Dell Inc.

DELL INC.

## Module summary

### Module summary—basic data analytics methods using R

Key points covered in this module:

- How to use basic analytics methods such as distributions, statistical tests and summary operations to investigate a data set
- How to use R to apply visualization patterns to better understand the data, help develop a model and derive hypotheses, and determine if our actions had a practical effect

# Advanced analytics—theory and methods

## Introduction



### Advanced analytics—theory and methods

Upon completing this module, you should be able to:

- ✓ Select an appropriate analytic technique based on the business problem reframed as an analytic challenge and based on the data's structure.
- ✓ Explain the technical foundations of commonly used analytic methods.
- ✓ Use R to fit, validate, and evaluate analytic models.

This module helps you understand the different analytic methods and learn where each analytic method is applicable.

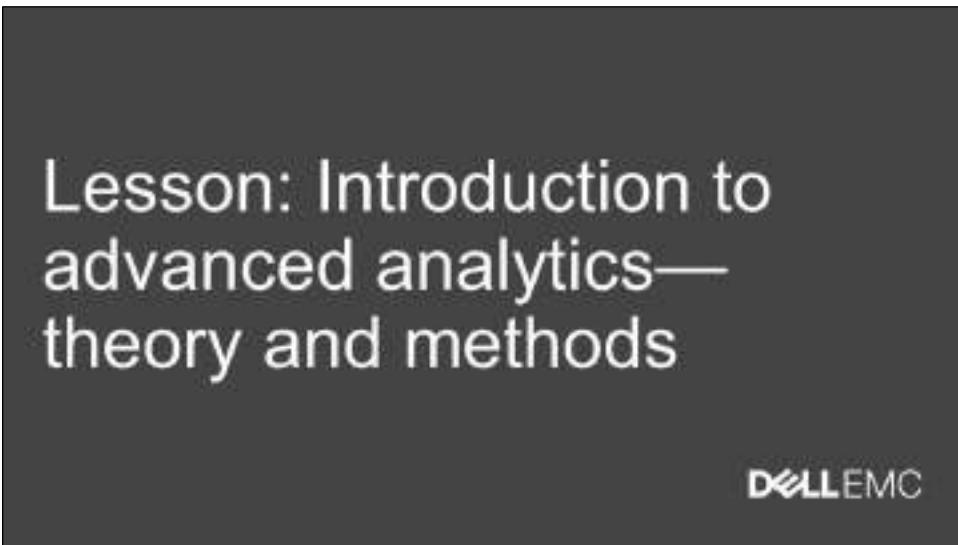
You can understand the technicality of each method and learn how to perform these solutions in R.

Upon completing this module, you should be able to:

- Select an appropriate analytic technique based on the business problem reframed as an analytic challenge and based on the data's structure.
- Explain the technical foundations of commonly used analytic methods.
- Use R to fit, validate, and evaluate analytic models.

## Lesson: Introduction to advanced analytics—theory and methods

Introduction



## Phase 3—model planning

**Model planning—key activities**

How do people generally solve this problem with the kind of data and resources I have?

- What are related or analogous problems? How are they solved? Can I do that? How can I improve on previous approaches?
- What are the model assumptions?
- Do I need extra data prep and transformations?

```
graph TD; Discovery --> DataPrep; DataPrep --> ModelPlanning; ModelPlanning --> ModelBuilding; ModelBuilding --> CommunicateResults; CommunicateResults --> Discovery; ModelPlanning --> ModelBuilding
```

DELL EMC

- In a typical data analytics project, you would have gone through:
  - Phase 1—Discovery—framing the business problem as an analytic challenge.
  - Phase 2—Data Preparation—exploring and conditioning the available data.
- Now, you must plan the model and determine the method to be used.

Model planning is the process of determining the appropriate analytic method based on the problem. It also depends on the type of data and the computational resources available.

## What kind of problem do I want to solve? How do I solve it?

| What kind of problem do I want to solve? How do I solve it?                       |                        |   |
|---|------------------------|---|
| Problem to solve  | Category of techniques | Covered in this course  |
| I want to group items by similarity.  | Clustering             | K-means clustering  |
| I want to find structure—commonalities, in the data.                              |                        |   |
| I want to discover relationships between actions or items.                        | Association rules      | Apriori   |
| I want to determine the relationship between the outcome and the input variables. | Regression             | Linear regression<br>Logistic regression                          |
| I want to analyze my text data.   | Text analysis          | Regular expressions, document representation—Bag of Words, TF-IDF |
| I want to assign known labels to objects.   | Classification         | Naïve Bayes<br>Decision trees                                     |
| I want to find the structure in a temporal process.                               |                        |   |
| I want to forecast the behavior of a temporal process.                            | Time series analysis   | ARIMA   |

This table lists the typical business questions, seen in column 1, addressed by a category of techniques or analytical methods, seen in column 2.

Some of the typical business questions for different categories of techniques are listed here:

- **Clustering**—How do I group these documents by topic? How do I group these images by similarity?
- **Association rules**—Based on historical data, what would someone buying a particular product also tend to buy?
- **Regression**—I want to predict the lifetime value of this customer. I want to predict the probability that this loan will default.
- **Classification**—Where in the catalog should I place this product? Is this email spam?
- **Time series analysis**—What is the likely future price of this stock? What will my sales volume be next month?
- **Text analysis**—Is this product review positive or negative?

There can be an overlap in the applicability of solutions. For example, consider the solution for questions such as "How do I group these documents?", "Is this email spam?", and "Is this product review positive or negative?"

## Lesson: Introduction to advanced analytics—theory and methods

These techniques can be used in combination to address analytic needs. For example, a project may start with text analysis, then do clustering on the document representations created by text analysis to discover document groupings and then lastly build a classification tool to more accurately classify new documents or provide insight on what drives the classification.

| Problem to solve  | Category of techniques | Covered in this course   |
|---|------------------------|--|
| I want to group items by similarity.<br>I want to find structure—commonalities in the data.                   | Clustering             | K-means clustering   |
| I want to discover relationships between actions or items.  | Association rules      | Apriori  |
| I want to determine the relationship between the outcome and the input variables.                             | Regression             | Linear regression<br>Logistic regression                             |
| I want to analyze my text data.   | Text analysis          | Regular expressions,<br>document representation—Bag of Words, TF-IDF |
| I want to assign known labels to objects.   | Classification         | Naïve Bayes<br>Decision trees  |
| I want to find the structure in a temporal process.<br>I want to forecast the behavior of a temporal process. | Time series analysis   | ARIMA  |

Similarly, more than one method can be used to solve the same problem. For example, time series analysis can be used to predict prices over time. Time series is used in cases where the past is observable to the participants, which is often true of stock and real estate. Sometimes, you can use regression methods, as well. However, regression is most effective when assigning effects to complicated patterns.

## Lesson: Introduction to advanced analytics—theory and methods

Column 3 in the table shown here lists the specific analytical methods that are detailed in the subsequent lessons in this module.

## Why these analytic techniques?

- Why these analytic techniques?
- Most popular, frequently used:
    - These techniques provide the foundation of data science skills on which to build
  - Relatively easy for new data scientists to understand and comprehend
  - Applicable to a broad range of problems in several verticals



The covered analytic techniques are commonly applied methods across several industry verticals. Also, these techniques provide a good foundation for understanding additional analytical methods.

## Lesson: K-means clustering

Introduction

Lesson: K-means  
clustering



## K-means clustering

### K-means clustering

During this lesson, the following topics are covered:

- Clustering—unsupervised learning method
- K-means clustering:
  - Use cases
  - The algorithm
  - Determining the optimum value for K
  - Diagnostics to evaluate the effectiveness of the method
  - Reasons to choose (+) and cautions (-) of the method

100

DATA SCIENCE

DELL INC.

This lesson covers K-means clustering.

## Clustering

### Clustering

- How do I group these documents by topic?
- How do I group my customers by purchase patterns?
- Sort items into groups by similarity:
  - Items in a cluster are more similar to each other than they are to items in other clusters.
  - Detail the properties that characterize similarity.
    - C1, detail the properties of distance, the "inverse" of similarity.
- Not a predictive method; finds similarities, relationships
- Example: K-means clustering



DATAVIZ

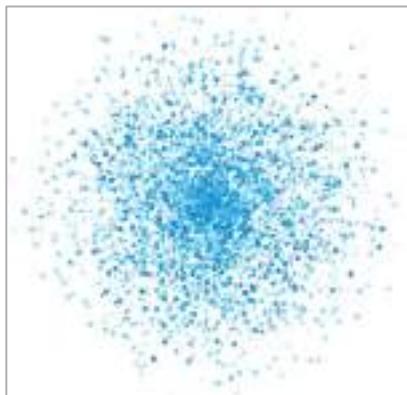
In machine learning, unsupervised refers to the problem of finding a hidden structure within unlabeled data. This lesson and the following lesson present two unsupervised learning methods: clustering and association rules.

**Clustering is a popular method used to form homogeneous groups within a dataset based on their internal structure.** Clustering is a method often used for exploratory analysis of the data. There are no "predictions" of any values completed with clustering, just finding the similarity between the data and grouping them into clusters.

The notion of similarities can be explained with the following examples.

Consider questions such as:

- How do I group these documents by topic?
- How do I perform customer segmentation to allow for targeted or special marketing programs?



**The definition of similarity is specific to the problem domain.** Similarity is defined as those data points with the same topic tag, or customers who can be profiled into a same age group, income, gender, or purchase pattern.

If you have a vector of measurements of an attribute of the data, the data points that are grouped into a cluster will have values for the measurement closer to each other than to those

data points grouped in a different cluster. In other words, the distance, an inverse of similarity, between the points within a cluster is always lower than the distance between points in a different cluster. **In a cluster, you end up with a tight, homogeneous group of data points that are far apart from those data points that end up in a different cluster.**

There are many clustering techniques. This lesson covers one of the most popular clustering methods, K-means clustering.

## K-means clustering—what is it?

### K-means clustering—what is it?

- Is a type of unsupervised learning used when you have unlabeled data
- Aims to partition  $n$  observations into  $k$  clusters in which each observation belongs to the cluster with the nearest mean
- Input: numerical. There must be a distance metric defined over the variable space.
  - Euclidian distance
- Output: the centers of each discovered cluster, and the assignment of each input datum to a cluster
  - Centroid

148 / 300 UNLABELED DATA

DELL EMC

K-means clustering is used to cluster numerical data.

Distance can be calculated in various ways, but four principles tend to hold true.

1. Distance is not negative; it is stated as an absolute value.
2. The distance from one point to itself is zero.
3. The distance from point I to point J is the same as the distance from point J to point I. Again, since the distance is stated as an absolute value, the starting and ending points can be reversed.
4. The distance between two points cannot be greater than the sum of the distance between each point and a third point.

**Euclidean distance** is the most popular method for calculating distance. Euclidian distance is an “ordinary” distance that one could measure with a ruler. In a single dimension, the Euclidian distance is the absolute value of the differences between two points, the straight-line distance between two points. In a plane with  $p_1$  at  $(x_1, y_1)$  and  $p_2$  at  $(x_2, y_2)$ , it is  $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ .

In  $N$  dimensions, the Euclidean distance between two points,  $p$  and  $q$ , is  $\sqrt{(\sum_{i=1}^N (p_i - q_i)^2)}$ , where  $p_i$  (or  $q_i$ ) is the coordinate of  $p$  (or  $q$ ) in dimension  $i$ .

Though there are many other distance measures, the Euclidian distance is the most commonly used distance measure and many packages use this measure.

The scale of the variables influences the Euclidian distance. Changing the scale—for example, from feet to inches—can significantly influence the results. Second, the equation ignores the relationship between variables. Lastly, the clustering algorithm is sensitive to outliers. If the data has outliers and removal of them is not possible, the results of the clustering can be substantially distorted.

**The centroid** is the center of the discovered cluster. K-means clustering provides this centroid as an output. When the number of clusters is fixed to  $k$ , *K-means clustering* gives a formal definition as an optimization problem: find the  $k$  cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

## K means clustering—use cases

### K means clustering—use cases

- Often an exploratory technique
  - Discover structure in the data
  - Summarize the properties of each cluster
- Sometimes a prelude to classification
  - Discovering the classes
- Examples
  - The height, weight, and average lifespan of animals
  - Household income, yearly purchase amount in dollars, number of household members of customer households
  - Patient record with measures of BMI, HBA1C, HDL
  - Cluster regions across a country based on sales, sensitivity, risk

100 / 300 pages of 1000 pages

DELL EMC

K-means clustering is often used as a lead-in to classification. It is primarily an exploratory technique to discover the structure of the data that you might not have noticed before; and, it is used as a prelude to more focused analysis or decision processes.

Some examples of the set of measurements based on which clustering can be performed are detailed in the slide.

In the patient record that includes measures such as BMI, HBA1C, and HDL, you could cluster patients into groups that define varying degrees of risk of a heart disease.

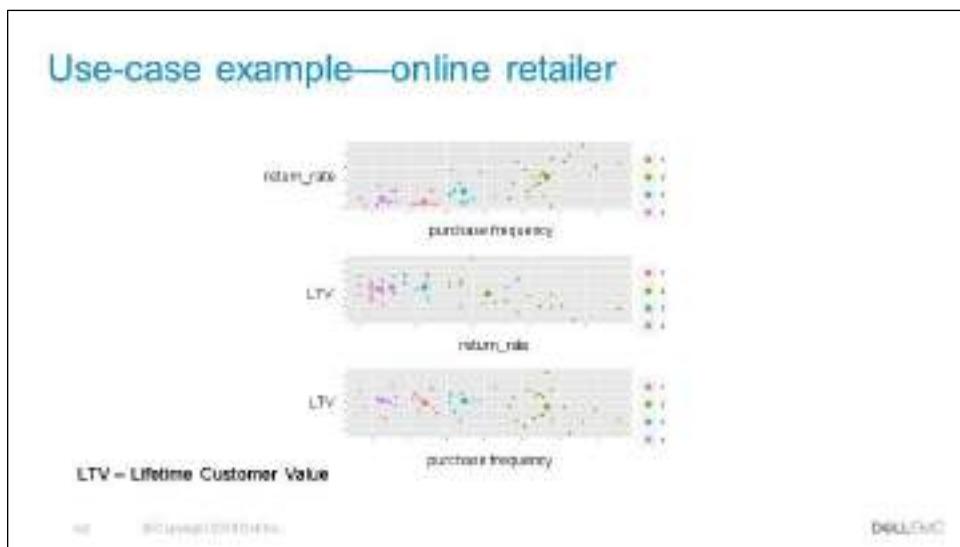
In classification techniques, the labels are known, whereas, in clustering, the labels are not known. Hence, clustering can be used to determine the structure in the data and summarize the properties of each cluster in terms of the measured centroids for the group. The clusters can define what the initial classes could be.

With a few dimensions, you can visualize the clusters. It gets harder to visualize as the dimensions increase.

There are many applications of the K-means clustering; examples include pattern recognition, image processing, machine vision, and so on.

In principle, you have several objects, and each object has several attributes. If you want to classify the objects based on the attributes, then you can apply this algorithm. For data scientists, K-means is an excellent tool to understand the structure of data and validate some of the assumptions provided by the domain experts pertaining to the data.

## Use-case example—online retailer



Here, you see an example of an online retailer. The unique selling point of this retailer is that they make the returns simple with an assumption that this policy encourages use and frequent customers are more valuable. So, you can validate this assumption.

Take a sample set of customers clustered on purchase frequency, return rate, and lifetime customer value (LTV).

Purchase frequency is defined as the number of visits a customer made that had a shopping cart transaction, in a month on average.

You can easily see that return rate has an important effect on customer value.

Cluster the customers into four groups, and then plot three graphs, taking two of the attributes in a graph. The data points are represented in the graphs by different colors for each cluster, and the larger dot represents the centroid for the group.

The groups can be defined broadly as follows:

- Group 1: Visit less frequently, low return rate, moderate LTV—ranked third
- Group 2: Visit often, return many of their purchases; lowest average LTV—counterintuitive

- Group 3: Visit often, return things moderately; high LTV—ranked second; happy medium
- Group 4: Visit rarely, do not return purchases; highest average LTV

Apparently, Group 3 is the ideal group—they visit often, return things moderately, and are high value.

The next questions are:

- Why is it that Group 3 is ideal?
- What are the people in these different groups buying?
- Is that affecting LTV?
- Can you raise the LTV of frequent customers, perhaps by lowering the cost of returns, or by somehow discouraging customers who return goods too frequently?
- Can you encourage the Group 4 customers to visit more, without lowering their LTV?
- Are more frequent customers more valuable?

You can see the range of questions that a data scientist can address with the initial analysis with K-means clustering.

## Algorithm

### Algorithm

Step 1: Choose K, then, select K random "centroids." In this example, K equals 3.

Step 2: Assign records to the cluster with the closest centroid.



DATA SOURCE: DELL INC.

Step 1: K-means clustering begins with the dataset segmented into K clusters.

Step 2: Observations are moved from cluster to cluster, to help reduce the distance from the observation to the cluster centroid.

## Algorithm, cont.

### Algorithm, cont.

Step 3: Recalculate the resulting centroids.  
Centroid: the mean value of all the records in the cluster.

Step 4: Repeat steps 2 and 3 until record assignments no longer change.



DATA SOURCE: DELL INC.

Step 3: When observations are moved to a new cluster, the centroid for the affected clusters must be recalculated.

Step 4: This movement and recalculation is repeated until movement no longer results in an improvement.

The model output is the final cluster centers and the final cluster assignments for the data.

Selecting the appropriate number of clusters, K, can be completed up front if you possess some knowledge on what the right number may be. Alternatively, you can try the exercise with different values for K and decide which clusters best suit your needs. Since it is rare that the appropriate number of clusters in a dataset is known, it is good practice to select a few values for K and compare the results.

The first partitioning should be accomplished with the same knowledge used to select the appropriate value of K—for example, domain knowledge about the market or industries.

If K was selected without external knowledge, the partitioning can be accomplished without any inputs.

After all observations are assigned to their closest cluster, the clusters can be

## Lesson: K-means clustering

evaluated for their in-cluster dispersion. Clusters with the smallest average distance are the most homogeneous. You can also examine the distance between clusters and decide if it makes sense to combine clusters that may be located close together. You can also use the distance between clusters to assess how successful the clustering exercise has been. Ideally, the clusters should not be located close together, as the clusters should be well separated.

## Picking K

### Picking K

Heuristic: find the elbow of the Within Sum of Squares (WSS) plot as a function of K.

$$WSS = \sum_{i=1}^k \sum_{j=1}^{n_i} \|x_{ij} - r_i\|^2$$

Where:  
 $k$ : # of clusters  
 $n_i$ : # points in  $i^{th}$  cluster  
 $r_i$ : centroid of  $i^{th}$  cluster  
 $x_{ij}$ :  $j^{th}$  data point of  $i^{th}$  cluster  
 $\|x_{ij} - r_i\|^2$ : distance between a centroid and point  $i$

Elbow method: look for the inflection point.

Scaling: it tries each variable to have equal impact in clustering.

DELL INC.

Practically based on the domain knowledge, a value for K is picked and the centroids are computed. Then, a different K is chosen, and the model is repeated to observe if it enhanced the cohesiveness of the data points within the cluster group. However, if there is no apparent structure in the data, you may have to try multiple values for K. It is an exploratory process.

Presented here is one of the heuristic approaches used for picking the optimal K for the given dataset. Within Sum of Squares (WSS) is a measure of how tight, on average, each cluster is. For  $k=1$ , WSS can be considered the overall dispersion of the data. WSS primarily is a measure of homogeneity. In general, more clusters result in tighter clusters. But, having too many clusters is overfitting.

The formula that defines WSS is shown. The graph depicts the value of WSS on the Y axis and the number of clusters on the X axis. The online retailer example data you reviewed earlier is the data with which the graph shown here is generated. The clustering for 12 different values is repeated. Going from one cluster to two, there is a significant drop in the value of WSS, since with two clusters you get more homogeneity. Look for the elbow of the curve, which provides the optimal number of clusters for the given data.

Visualizing the data helps in confirming the optimal number of clusters. Reviewing the three pair-wise graphs plotted for the online retailer example earlier, you can see that having four groups sufficiently explained the data; and, from the graph,

## Lesson: K-means clustering

you can also see that the elbow of the curve is at 4.

Also, another important concept here is scaling. Two variables with two different units—for example, one variable in millions and another in hundreds—have a different impact on clustering. The higher the value of the variable, the higher the impact on clustering, which skews the results. Scaling helps ensure that the impact of each variable for clustering is equal.

## Diagnostics—evaluating model

### Diagnostics—evaluating model

- Do the clusters look separated in at least some of the plots when you do pair-wise plots of the clusters?
  - Pair-wise plots can be used when there are not many variables.
- Do you have any clusters with few data points?
  - Try decreasing the value of K.
- Are there splits on variables that you would expect but do not see?
  - Try increasing the value K.
- Do any of the centroids seem too close to each other?
  - Try decreasing the value of K.
- Do the means of variables used for clustering vary across the clusters?
  - Try decreasing the value of K.

00 00000000000000000000000000000000

DELL INC.

How do you know that you have good clusters?

Pair-wise plots of the clusters provide a good visual confirmation that the clusters are homogeneous. When the dimensions of the data are not significantly large, this method helps in determining the optimal number of clusters. With these plots, you should be able to determine if the clusters look separated in at least some of the plots. They will not be separated in all of the plots. This effect can be seen even with the online retailer example you saw earlier. Some of the clusters get mixed in together, in some dimensions.

If you feel that your clusters are too small, it indicates that you have a large value for K and that K must be reduced; try a smaller K. It may be the outliers in the data that tend to cluster into clusters with less data points.

Alternatively, if you see that there are splits that you expected but are not seen in the clusters—for example, you expect two different income groups and you do not see them—you should try a bigger value for K.

If the centroids seem too close to each other, then you should try decreasing the value of K.

## K-means clustering—reasons to choose (+) and cautions (-)

| K-means clustering—reasons to choose (+) and cautions (-)                            |   |
|--|---|
| Reasons to choose (+)  | Cautions (-)  |
| Easy to implement  | Does not handle categorical variables   |
| Easy to assign new data to existing clusters<br>Which is the nearest cluster center? | Sensitive to initialization—first guess   |
| Concise output<br>Coordinates the K cluster centers                                  | Variables should all be measured on similar or compatible scales<br>Not scale-invariant   |
|  | K, the number of clusters, must be known or decided in a way based on theoretical deduction<br>Wrong guess: possible poor results |
|  | Tends to produce "round," equal-sized clusters<br>Not always desirable  |

K-means clustering is easy to implement, and it produces concise output. It is easy to assign new data to the existing clusters by determining which centroid the new data point is closest to it.

However, K-means works only on the numerical data and does not handle categorical variables. It is sensitive to the initial guess on the centroids. It is important that the variables must be all measured on similar or compatible scales. If you measure the living space of a house in square feet, and the cost of the house in thousands of dollars—that is, one unit is \$1000—and then you change the cost of the house to dollars, so one unit is \$1, then the clusters may change. **K should be decided ahead of the modeling process.** Wrong guesses for K may lead to improper clustering.

K-means tends to produce rounded and equal sized clusters. If you have clusters that are elongated or crescent shaped, then K-means may not be able to find these clusters appropriately. The data, in this case, may have to be transformed before modeling.

| Reasons to choose (+)  | Cautions (-)  |
|--|---|
| Easy to implement  | Does not handle categorical variables   |
| Easy to assign new data to existing clusters<br>Which is the nearest cluster center? | Sensitive to initialization—first guess   |
| Concise output<br>Coordinates the K cluster centers                                  | Variables should all be measured on similar or compatible scales<br>Not scale-invariant!  |
|  | K, the number of clusters, must be known or decided in a way based on theoretical deduction<br>Wrong guess: possibly poor results |
|  | Tends to produce "round," equal-sized clusters<br>Not always desirable  |

Some of the cautions explained here can be overcome to some extent through other techniques.

- Sensitive to Initialization: Can be adjusted with nstart and number of iterations.
- Scale Invariant: All the variables can be run through scaling technique for equal impact.
- Number of clusters: New methodologies have been identified that would enable you to know the optimum number of clusters.
- Round equal-sized clusters: Can be changed by changing the number of clusters.

## Check your knowledge

### Check your knowledge

1. Why is K-means clustering considered an unsupervised machine learning algorithm?
2. Detail the four steps in the K-means clustering algorithm.
3. How do you use WSS to pick the value of K?
4. What is the most a common measure of distance used with K-means clustering algorithms?
5. The attributes of a dataset are purchase decision (Yes/No), gender (M/F), income group (<10K, 10-50K, >50K). Can you use K-means to cluster this dataset?



000 00000000000000000000000000000000

Dell EMC

## Check your knowledge



### Discussion

## Question / Discussion Topic:

1. Why is K-means clustering considered an unsupervised machine learning algorithm?
2. Detail the four steps in the K-means clustering algorithm.
3. How do you use WSS to pick the value of K?
4. What is the most a common measure of distance used with K-means clustering algorithms?
5. The attributes of a dataset are purchase decision (Yes/No), gender (M/F), income group (<10K, 10-50K, >50K). Can you use K-means to cluster this dataset?

**Discussion Notes:**

## K-means clustering—summary

### K-means clustering—summary

During this lesson, the following topics were covered:

- Clustering—unsupervised learning method
- K-means clustering
- Use cases with K-means clustering
- The K-means clustering algorithm
- Determining the optimum value for K
- Diagnostics to evaluate the effectiveness of K-means clustering
- Reasons to choose (+) and cautions (-) of K-means clustering



The summary of key topics presented in this lesson is listed here. Take a moment to review these topics.

## Lesson: Association rules

Introduction

# Lesson: Association rules

DELL EMC

## Association rules

### Association rules

During this lesson, the following topics are covered:

- Association rules mining
- Apriori algorithm
- Prominent use cases of association rules
- Support and confidence parameters
- Lift and leverage
- Diagnostics to evaluate the effectiveness of rules generated
- Reasons to choose (+) and cautions (-) of the Apriori algorithm

The topics covered in this lesson are listed.

## Grocery store—scenario

### Grocery store—scenario

- In order to increase the volume of sales, grocery store managers may want to perform an analysis for understanding the products that shoppers purchase together.
- For example, If the managers find that there is a higher probability of a customer purchasing bread and milk together, they may want those items in aisles that are close to each other. This way, there is more possibility of customers buying milk easily when they come in to buy bread, which increases the volume of sales.
- At this point, association rules come into play and help identify those patterns.



Here, you see a grocery store scenario, for an example. So they can increase the volume of sales, grocery store managers may want to perform an analysis for understanding the products that shoppers visit together, so they can place them closer in the aisles. This strategy increases the probability of greater purchases revenue.

At this point, association rules come into play.

### Discussion Notes:

## Association rules

### Association rules

- Help identify interesting patterns and connections among sets of items
  - Rules take the form of "If X is observed, then Y is also observed"
  - The definition of "interesting" varies with the algorithm used for discovery.
- Use case: Understand customer buying habits by finding associations between the different items that customers place in their "shopping basket"
  - Known as market basket analysis
- Not a predictive method

000

DATA SCIENCE

DELL EMC

Association rules is another unsupervised learning method. There is no prediction performed, but this method is used to discover relationships within the data.

The example questions are:

- Which of my products tend to be purchased together?
- What will other people who are like this person or product tend to buy or watch or click for other products I may offer?

In the online retailer example you analyzed in the previous lesson, you could use association rules to discover what products are purchased together within the group that yielded maximum LTV. For example, if you set up the data appropriately, you could explore to further discover which products people in GP4 tend to buy together and derive any logical reasons for high rate of returns. You can discover the profile of purchases for people in different groups—for example, people who buy high-heeled shoes and expensive purses tend to be in GP4 or people who buy walking shoes and camping gear tend to be in GP2, and so on.

The goal with association rules is to discover interesting relationships among the variables, and the definition of interesting depends on the algorithm used for the discovery.

The **rules** you discover are of the form that "When I observe X, I also tend to

observe Y."

Examples of interesting relationships are those rules identified with a measure of confidence—with a value greater than or equal to a predefined threshold—with which a rule can be stated based on the data.

## Association rules—examples

### Association rules—examples

Examples where association rules could be applied:



Amazon: Notice the "Customers Who Bought This Item Also Bought" section in the Amazon website.

Netflix: For every movie watched, there is a recommendation for movie Y "Because you watched movie X."

YouTube: Based on your viewing pattern, YouTube has a "Recommended" section that finds the relationships.

Source: <http://www.cs.cmu.edu/~mlittman/ML-Notes/ML-Notes-10.pdf>

Dell EMC

Here are few examples where association rules are used.

## Algorithm for association rules—Apriori

### Algorithm for association rules—Apriori

- Apriori—an algorithm for mining frequent itemsets for the Boolean association rules
  - Uses a bottom-up approach where frequent subsets are extended one item at a time
  - Designed to operate on datasets containing transactions
- Used over itemsets—sets of discrete variables that are linked
- Possible transactional datasets
  - Retail items that are purchased together
  - A set of tasks completed in one day
  - A set of links one user clicks in a single session
- Four common ways to measure association
  - Support
  - Confidence
  - Lift
  - Leverage

DATA SCIENCE

DELL EMC

Association rules are designed for in-database mining over transactions in databases. The algorithm uses a bottom-up approach—it starts with one item that has most frequently occurred in all the baskets. Then, the next item that is most frequently associated with the first item creates the new subset. The process is repeated for every item in the basket that would satisfy the conditions the user gives in the Apriori algorithm.

Association rules are used over transactions that consist of itemsets. Itemsets are discrete sets of items that are linked together. For example, they could be a set of retail items purchased together in one transaction. Association rules are sometimes called **market basket analysis**, and you can think of an itemset as everything in your shopping basket. Market basket analysis provides insights into which products tend to be purchased together and which are most amenable to promotion.

You can also group the tasks completed in one day or set of links a user clicks in a single session into a basket or an itemset, for discovering associations.

Apriori is one of the earliest and the most commonly used algorithms for association rules. It is the focus for the rest of this lesson.

The four common ways to measure association are support, confidence, lift, and leverage, to be discussed in detail in the next few slides.

## Apriori algorithm

### Apriori algorithm

- Earliest of the association rule algorithms
- Frequent itemset: a set of items L that appear together often enough:
  - Formally: meets a minimum support criterion
  - Support: the percentage of transactions that contain the itemset, which shows how popular an itemset is; this percentage is measured by the proportion of transactions that contain it
- Apriori property: Any subset of a frequent itemset is also frequent
  - It has at least the support of its superset.

34

DATA SCIENCE

DELL INC.

The Apriori algorithm uses the notion of frequent itemset. As the name implies, the frequent itemsets are a set of items "L" that appear together often enough. The term "often enough" is formally defined with a support criterion, where the support is defined as the percentage of transactions that contain "L".

For example, imagine L is defined as an itemset {shoes, purses}, and support is defined as 50 percent. If 50 percent of the transactions have this itemset, then the L is a frequent itemset. It is apparent that if 50 percent of itemsets have {shoes, purses} in them, then at least 50 percent of the transactions have either {shoes} or {purses} in them. This relationship is an **Apriori property**, which states that **any subset of a frequent itemset is also frequent**. The Apriori property provides the basis for the Apriori algorithm detailed in the subsequent slides.

Any itemset that does not occur often enough will not occur more often by adding another item. This greatly reduces the number of itemsets that need to be checked. Otherwise, all non-empty  $2^n - 1$  subsets would need to be examined against the minimum support criterion.

## Apriori algorithm—support example

### Apriori algorithm—support example

Consider the grocery store example with 1000 transactions and the following items in the basket. Minimum support is 50 percent.

| ITEMSET     | SUPPORT |
|-------------|---------|
| Milk        | 70%     |
| Bread       | 61.2%   |
| Milk, Bread | 62.7%   |

itemset {Milk, Bread} has minimum support  
Possible rules are Milk → Bread, Bread → Milk

If you find that an itemset greater than a certain proportion tends to have significant impact on profits, then that proportion can be used as threshold for support. You can identify itemsets with support values greater than the threshold as significant itemsets.

DATA SCIENCE

Here is a closer look at the Apriori algorithm.

The Apriori algorithm uses the notion of frequent itemset. Support explains how popular an itemset is as measured by the proportion of transactions that contain this itemset.

In this example, you have a set of artificially created transaction records detailing grocery store transactions. Here, you find records in which Milk, Bread, Eggs have a support of over 50 percent.

As the itemset {Milk, Bread} has a minimum support of over 50 percent, you can state the following rules:

- Milk → Bread
- Bread → Milk

## Apriori algorithm—confidence

### Apriori algorithm—confidence

- Iteratively grow the frequent itemsets from size 1 to size K, or until you run out of support.
  - Apriori property tells you how to prune the search space.
- Frequent itemsets are used to find rules  $X \rightarrow Y$  with a minimum **confidence**.
  - **Confidence:** The percentage of transactions that contain X that also contain Y.
- Output: The set of all rules  $X \rightarrow Y$  with minimum support and confidence.

10

DATA SCIENCE

DELL EMC

As mentioned previously, Apriori is a bottom-up approach where you start with all the frequent itemsets of size 1 first—for example, shoes, purses, hats, and so on—and determine the support. Then, you start pairing them. You find the support for, say, {shoes, purses} or {shoes, hats} or {purses, hats}.

Suppose that you set your threshold as 50 percent; you find those itemsets that appear in 50 percent of all transactions. You scan all the itemsets and prune away those itemsets that have less than 50 percent support—they appear in less than 50 percent of the transactions—and keep the ones that have sufficient support. The word "prune" is used as it would be in gardening, where you prune away the excess branches of your bushes.

The Apriori property provides the basis to prune over the transactions—search space—and to stop searching further if the support threshold criterion is **not** met. If the support criterion is met, then you grow the itemset and repeat the process until you have the specified number of items in an itemset or you run out of support.

You now use the frequent itemsets to find your rules, such as X implies Y. Confidence is the percent of transactions that contain X that also contain Y. For example, if you have frequent itemset {milk, bread, cookies} and consider subsets {milk, bread}, and if 80 percent of the transactions that have {milk, bread} also have {cookies}, then you define confidence for the rule that {milk, bread} implies {cookies} as 80 percent.

The outputs of the Apriori are the rules with minimum support and confidence.

## Apriori algorithm—confidence example

### 030. Apriori algorithm—confidence example

In the grocery store example, you have 1000 records with the following combinations:

|       | Milk | Bread | Total |
|-------|------|-------|-------|
| Milk  | 44   | 186   | 230   |
| Bread | 24   | 627   | 651   |
| Total | 78   | 813   | 990   |

Out of 813 shoppers who buy bread, 627 buy Milk, as well.

$$\text{Milk} \rightarrow \text{Bread} = 627/813 = 77.07\%$$

$$\text{Bread} \rightarrow \text{Milk} = 627/813 = 77.12\%$$

One drawback of the confidence measure is that it might misrepresent the importance of an association. It only accounts for how popular Milk is, but not Bread. If Bread is also popular, in general, there is a higher chance that a transaction containing Milk also contains Bread, inflating confidence measure.

To account for the base popularity of both constituent items, you use a third measure called lift.

The drawback of the confidence measure is misrepresentation of the importance of association. Consider the milk and bread example. Here, you see that the confidence of buying bread after milk is 77.1 percent. But, if there is a scenario where the support for bread is already at 80 percent, then there is a high probability that the resulting confidence is because of the high support for bread. So, along with the confidence, you must look into lift that also takes into account the popularity of both items.

## Lift and leverage

### Lift and leverage

Lift explains how likely item Y is purchased when item X is purchased, as if X and Y are statistically independent:

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X) * \text{Support}(Y)}$$

$\text{Lift} = 1$  indicates no association between items. The relationship is coincidental.  
 $\text{Lift} > 1$  indicates that item Y is likely to be bought if item X is bought. Relationship is interesting.  
 $\text{Lift} < 1$  indicates that item Y is unlikely to be bought if item X is bought.

Leverage measures the difference in the probability of X and Y appearing together in the dataset compared to what would be expected if X and Y were statistically independent:

$$\text{Leverage}(X \rightarrow Y) = \text{Support}(X \cup Y) - \text{Support}(X) * \text{Support}(Y)$$

The common measures that the Apriori algorithm uses are **support** and **confidence**. Rank all the rules based on the support and confidence, and filter out the most interesting rules.

There are other measures to evaluate candidate rules; **lift** and **leverage** are two such measures.

Lift measures how many times more often X and Y occur together than expected if they were statistically independent. It is a measure of how X and Y are related rather than coincidentally happening together.

Leverage is a similar notion; but, instead of a ratio, it is the difference.

Leverage measures the difference in the probability of X and Y appearing together in the dataset compared to what would be expected if X and Y were statistically independent.

For more measures, visit this web page:  
[michael.hahsler.net/research/association\\_rules/measures.html](http://michael.hahsler.net/research/association_rules/measures.html)

## Apriori algorithm—lift and leverage example

**Apriori algorithm—lift and leverage example**

In this grocery store example:

|       | Milk | Bread | Total |
|-------|------|-------|-------|
| Milk  | 41   | 186   | 227   |
| Bread | 24   | 627   | 651   |
| Total | 65   | 813   | 768   |

You calculated the confidence for the following rules:  
Milk → Bread =  $627/768 = 89.52\%$   
Bread → Milk =  $627/813 = 77.12\%$

Lift of these two rules =  $0.627/0.700 \cdot 0.813 = 1.10$ .  
Since lift is  $> 1$ , the relationship is interesting and true.

140 800-1000-1100-1200  
Dell EMC

The lift is the ratio of [the probability of buying milk with bread], divided by [the probability of buying milk multiplied by the probability of buying bread], which is 0.627 divided by (0.700 x 0.813), which equals 1.10.

The lift being close to the value of 1 indicates that the rule is coincidental. With larger values of lift, you may say that the rule is true and not coincidental.

## Sketch of algorithm

### Sketch of algorithm

- If  $L_k$  is the set of frequent k-itemsets:
  - Generate the candidate set  $C_{k+1}$  by joining  $L_k$  to itself.
  - Prune out the  $(k+1)$ -itemsets that do not have minimum support. Now, you have  $L_{k+1}$ .
- You know this algorithm catches all the frequent  $(k+1)$ -itemsets by the Apriori property.
  - A  $(k+1)$ -itemset cannot be frequent if any of its subsets are not frequent.
- Continue until you reach  $k_{max}$  or run out of support.
- From the union of all the  $L_k$ , find all the rules with minimum confidence.

1/1

DATA SCIENCE

DELL INC.

This sketch gives a more formal definition of the Apriori algorithm.

Step 1 is identifying the frequent itemsets by starting with each item on the transactions that meet the support level. Then, you grow each itemset joining another itemset and determine the support for the new grown itemset.

Prune all the itemsets that do not meet the minimum support.

Repeat the growing and pruning until you reach the specified number of items in an itemset or you run out of support.

Then, form rules with the union of all the itemsets that you retained that meet the minimum confidence threshold.

Returning to the grocery store example allows for a better understanding of this algorithm.

## Step 1—1-itemsets (L1)

- Let min\_support = 0.5
- 1000 transactions
- Scan the database
- Prune

| Frequent Itemset | Count |
|------------------|-------|
| Bread            | 813   |
| Milk             | 700   |
| Bacon            | 631   |
| Potatoes         | 550   |

The first step is to start with a 1-element itemset and let the support be 0.5. You scan the database and count the occurrences of each attribute.

The itemsets that meet the support criteria are the ones that are not pruned, or struck off.

## Step 2—2-itemsets (L2)

### Step 2—2-itemsets (L2)

- Join L1 to itself
- Scan the database to get the counts
- Prune

| ITEMSET    | SUPPORT |
|------------|---------|
| Milk,Soda  | 544     |
| Milk,Bread | 621     |

The itemsets that you end up with at step 1 are {Milk}, {Potatoes}, {Bread}, and {Soda}.

In step 2, you join—or, grow—these itemsets with two elements in each itemset as {Milk, Potatoes}, {Milk, Bread}, {Milk, Soda}, {Potatoes, Soda}, {Potatoes, Bread}, and {Soda, Bread} and determine the support for each of these combinations.

What survive the pruning are {Milk, Soda} and {Milk, Bread}.

## Step 3—3-itemsets

### Step 3—3-itemsets

- You have run out of support.
- Candidate rules come from L2:
  - Milk → Soda
  - Soda → Milk
  - Milk → Bread
  - Bread → Milk

140

DATA SCIENCE

DELL EMC

When you grow the itemsets to 3, you run out of support.

At this point, you stop and generate rules with results in step 2.

The rules that come from step 2 are shown.

Obviously, depending on what you are trying to do—for example, predict who will buy Milk—some rules are more useful than others, independent of confidence.

## Finally—find confidence rules

| Rule               | Set   | Count | Conf.            | Confidence      |
|--------------------|-------|-------|------------------|-----------------|
| If Milk THEN Soda  | Milk  | 718   | Milk AND Soda    | 544 / 718 = 76% |
| If Milk THEN Bread | Milk  | 718   | (Milk AND Bread) | 521 / 718 = 73% |
| If Soda THEN Milk  | Soda  | 633   | Soda AND Milk    | 544 / 633 = 86% |
| If Bread THEN Milk | Bread | 618   | (Bread AND Milk) | 621 / 618 = 77% |

If you want confidence > 80%:  
If Soda THEN Milk

100 - 80 = 20%  
80 / 100 = 80%

DELL EMC

After you have the rules, compute the confidence for each rule. The table lists the rules and the computation of confidence.

You see here that Soda → Milk has an 86 percent confidence.

## Diagnostics

### Diagnostics

- Do the rules make sense?
  - What does the domain expert say?
- Make a "test set" from hold-out data:
  - Enter some market baskets with a few items missing, selected at random. Can the rules determine the missing items?
  - Remember, some of the test data may not cause a rule to fire.
- Evaluate the rules by lift or leverage.
  - Some associations may be coincidental, or obvious.

100

DATA SCIENCE

DELL EMC

The first check on the output is to determine if the rules make any sense. The domain expertise provides inputs for this check.

In the example of the grocery store, you had 1000 transactions that you worked with for the discovery of rules. Now, assume that you had 1500 transactions. You can randomly select 500 transactions out of this total and keep it aside as hold-out data, and then run the discovery of rules on the remaining 1000 transactions. The 500 records you kept aside are known as the **hold-out data**.

You can use the data as a test set and drop some items from the transactions randomly. When you run the association rules again on the test set, determine if the algorithm predicts the missing data or the items dropped. It should be noted that some of the test data may not cause the rule to fire.

It is important to evaluate the rules with lift or leverage. When you mine data with association rules, several rules are generated that are purely coincidental.

If 95 percent of your customers buy X, and 90 percent of customers buy Y, then X and Y occur together 85 percent of the time, even if there is no relationship between the two. The measure of lift ensures that interesting rules are identified rather than coincidental ones.

## Apriori—reasons to choose (+) and cautions (-)

| Apriori—reasons to choose (+) and cautions (-)  |  |
|---|--|
| Reasons to choose (+)   | Cautions (-)   |
| Easy to implement <ul style="list-style-type: none"><li>+ Uses a clever observation to prune the search space<ul style="list-style-type: none"><li>- Apriori property</li></ul></li></ul> | Requires many database scans <ul style="list-style-type: none"><li>- Exponential time complexity</li></ul>   |
| Easy to parallelize   | <ul style="list-style-type: none"><li>+ Can mistakenly find spurious, or coincidental relationships<ul style="list-style-type: none"><li>- Addressed with lift and leverage measures</li></ul></li></ul> |

While the Apriori algorithm is easy to implement and parallelize, it is computationally expensive. One of the major drawbacks with the algorithm is that many spurious rules tend to get generated that are practically not useful. These spurious rules are generated due to coincidental relationships between the variables. Lift and Leverage measures must be used to prune out these rules.

## Check your knowledge

### Check your knowledge

1. What is the Apriori property, and how is it used in the Apriori algorithm?
2. List three popular use cases of the association rules mining algorithms.
3. What is the difference between lift and leverage? How is lift used in evaluating the quality of rules discovered?
4. Define support and confidence.
5. How do you use a hold-out dataset to evaluate the effectiveness of the rules generated?



## Check your knowledge



### Discussion

## Question / Discussion Topic:

1. What is the Apriori property, and how is it used in the Apriori algorithm?
2. List three popular use cases of the association rules mining algorithms.
3. What is the difference between lift and leverage? How is lift used in evaluating the quality of rules discovered?
4. Define support and confidence.
5. How do you use a hold-out dataset to evaluate the effectiveness of the rules generated?

## Discussion Notes:

## Association rules—summary

### Association rules—summary

During this lesson, the following topics were covered:

- Association rules mining
- Apriori algorithm
- Prominent use cases of association rules
- Support and confidence parameters
- Lift and leverage
- Diagnostics to evaluate the effectiveness of rules generated
- Reasons to choose (+) and cautions (-) of the Apriori algorithm



DATA SCIENCE

DELL EMC

This lesson covered these topics. Take a moment to review them.

## Lesson: Linear regression

Introduction

Lesson: Linear regression

DELL EMC

## Linear regression lesson topics

### Linear regression lesson topics

During this lesson, the following topics are covered:

- Technical description of a linear regression model
- Common use cases for the linear regression model
- Interpretation and scoring with the linear regression model
- Diagnostics for validating the linear regression model
- The reasons to choose (+) and cautions (-) of the linear regression model

The topics covered in this lesson are listed.

## Regression

### Regression

- Regression focuses on the relationship between an outcome and its input variables.
  - Provides an estimate of the outcome based on the input values
  - Models how changes in the input variables affect the outcome
- The outcome can be continuous or discrete.
- Possible use cases:
  - Estimate the lifetime value (LTV) of a customer and understand what influences LTV.
  - Estimate the probability that a loan will default and understand what leads to default.
- **Approaches: linear regression and logistic regression**

100

DATA SCIENCE

DELL EMC

Francis Galton coined the term regression in the 19th century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards an average—a phenomenon also known as regression toward the mean.

Specifically, regression analysis helps one understand how the value of the dependent variable—also known as outcome—changes when any one of the independent, or input, variables changes, while the other independent variables are held fixed. Regression analysis estimates the conditional expectation of the outcome variable given the input variables—that is, the mean value of the outcome variable when the input variables are held fixed.

Regression focuses on the relationship between the outcome and the inputs. It also provides a model that has some **explanatory value**, in addition to estimating outcomes. Although social scientists use regression primarily for its explanatory value, data scientists apply regression techniques as predictors or classifiers.

The outcome can be continuous or discrete. For continuous outcomes, such as income, this lesson examines the use of **linear regression**. For discrete outcomes of a categorical attribute, such as success or fail, gender, or political party affiliation, the next lesson presents the use of **logistic regression**.

## Linear regression

### Linear regression

- Used to estimate a continuous value as a linear, additive, function of other variables:
  - Income as a function of years of education, age, and gender
  - House sales price as function of square footage, number of bedrooms and bathrooms, and lot size
- Outcome variable is continuous.
- Input variables can be continuous or discrete.
- Model output:
  - A set of estimated coefficients that indicate the relative impact of each input variable on the outcome
  - A linear expression for estimating the outcome as a function of input variables

© 2018 Dell Inc. All rights reserved.

DELL.COM

Linear regression is a commonly used technique for modeling a continuous outcome. It is simple and works well in many instances. Dell recommends that linear regression should be tried. And, if it is determined that the results are not reliable, other more complicated models should be considered. Alternative modeling approaches include ridge regression, local linear regression, regression trees, and neural nets—these models are out of scope for this course.

Linear regression models a continuous outcome, such as income or housing sales prices, as a linear or additive function of other input variables. The input variables can be continuous or discrete.

## Linear regression model

### Linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \epsilon$$

where

$y$  is the outcome variable

$x_j$  are the input variables, for  $j = 1, 2, \dots, p-1$

$\beta_0$  is the value of  $y$  when each  $x_j$  equals zero

$\beta_j$  is the change in  $y$  based on a unit change in  $x_j$

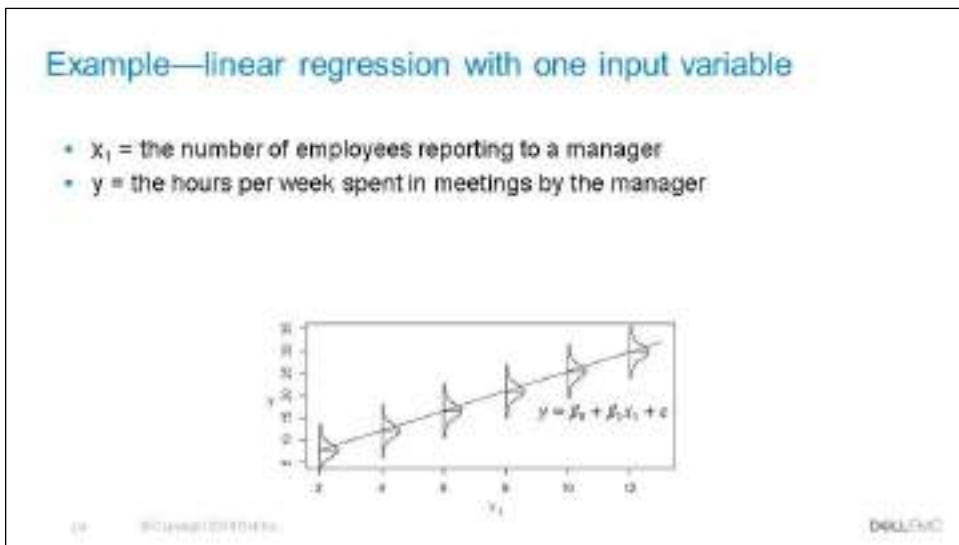
$\epsilon \sim N(0, \sigma^2)$  and the  $\epsilon$ 's are independent of each other

DATA SCIENCE

DELL INC.

In linear regression, the outcome variable is expressed as a linear combination of the input variables. For a given set of input variables, the linear regression model provides the expected outcome value. Unless the situation being modeled is purely deterministic, there will be some random variability in the outcome. This random error, denoted by  $\epsilon$ , is assumed to be normally distributed with a mean of zero and a constant variance ( $\sigma^2$ ).

## Example—linear regression with one input variable



In this example, the human resources department decides to examine the effect that the number of employees directly reporting to a manager has on how many hours per week the manager spends in meetings. The expected time spent in meetings is represented by the equation of a line with unknown intercept and slope. Suppose that the true value of the intercept is 3.27 hours and the true value of the slope is 2.2 hours per employee. Then, a manager can expect to spend an extra 2.2 hours per week in meetings for every additional employee.

The rotated normal distribution plots provided at specific values of  $x_1$  represent the distribution of the error term. For example, a typical manager with eight employees may be expected to spend 20.87 hours per week in meetings, but any amount of time from 15 to 27 hours per week is probable.

This example illustrates a theoretical regression model. In practice, it is necessary to collect and prepare the necessary data and use a software package such as R to estimate the values of the coefficients. Coefficient estimation is covered later in this lesson.

More variables could be included in this model. For example, a categorical attribute can be added to this linear regression model to account for the manager's functional organization, such as engineering, finance, manufacturing, or sales. It may be tempting to include one variable,  $x_2$ , to represent the organization and denote engineering by 1, finance by 2, manufacturing by 3, and sales by 4.

## Lesson: Linear regression

However, such an approach incorrectly suggests that the interval between the assigned numeric values has meaning—for example, sales is three more than engineering. The proper implementation of categorical attributes in a regression model will be addressed next.

## Representing categorical attributes

**Representing categorical attributes**

For a categorical attribute with  $m$  possible values:

- Add  $m-1$  binary (0/1) variables to the regression model.
- The remaining category is represented by setting the  $m-1$  binary variables equal to zero.

$$y = \beta_0 + \beta_1 \text{employees} + \beta_2 \text{finance} + \beta_3 \text{mfg} + \beta_4 \text{sales} + \epsilon$$

| POSITION                                | FUNCTION      | NUMBER OF EMPLOYEES |
|---|---------------|---------------------|
| Finance manager with 6 employees        | Finance       | 011000              |
| Manufacturing manager with 10 employees | Manufacturing | 100,00              |
| Sales manager with 10 employees         | Sales         | 000,11              |
| Engineering manager with 8 employees    | Engineering   | 000000              |

DELL INC.

In expanding the previous example to include the manager's functional organization, the input variables, denoted earlier by the  $x$ 's, have been replaced by more meaningful variable names. In addition to the *employees* variable for the number of employees reporting to a manager, three binary variables have been added to the model to identify finance, manufacturing (*mfg*), and sales managers. If a manager belongs to either of these functional organizations, the corresponding variable is set to 1. Otherwise, the variable is set to 0. Thus, for four functional organizations, engineering is represented by the case where the three binary variables are each set to 0. For this categorical variable, engineering is considered the reference level.

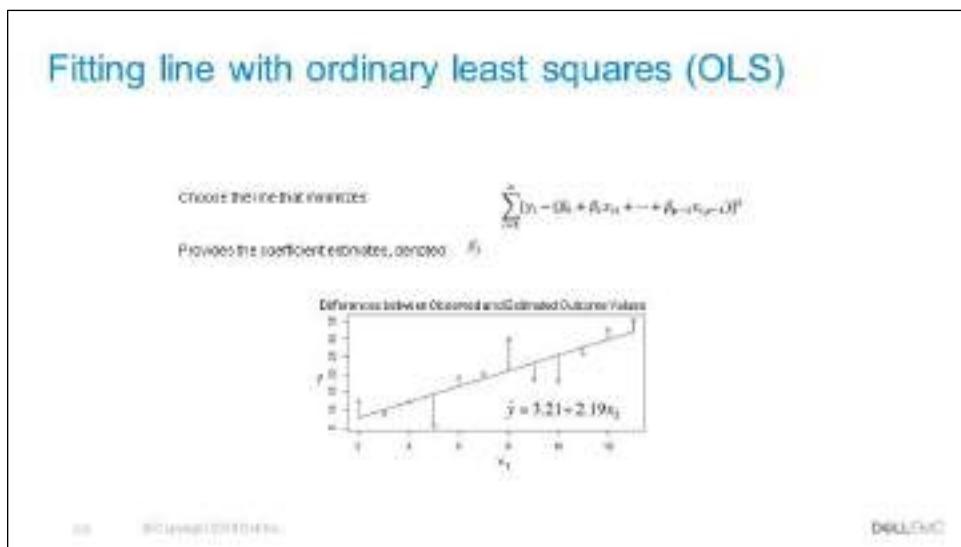
For example, the coefficient of finance denotes the relative difference from the reference level. Choosing a different organization as the reference level changes the coefficient values, but not their relative differences. Interpreting the coefficients for categorical variables relative to the reference level is covered later in this lesson.

In general, a categorical attribute with  $m$  possible distinct values can be represented in the linear regression model by adding  $m-1$  binary variables. For a categorical attribute, such as gender with only two possible values, female or male, then only one binary variable must be added with one value assigned a 0 and the other value assigned a 1.

## Lesson: Linear regression

Suppose it was decided to include the manager's U.S. state of employment in the regression model. Then, 49 binary variables must be added to the regression model to account for 50 states. However, that many categorical values can be cumbersome to interpret or analyze. Alternatively, it may make more sense to group the states into geographic regions or into other groupings such as type of location: headquarters, plant, field office, or remote. In the latter case, only three binary variables must be added.

## Fitting line with ordinary least squares (OLS)



After a dataset has been collected, the objective is fit the “best” line to the data points. A common approach to determine the best-fitting line is to choose the line that minimizes the sum of the squares of the differences between the observed outcomes in the dataset and the estimated outcomes based on the equation of the fitted line. This method is known as Ordinary Least Squares (OLS). If there is one input variable, the differences or distances between the observed outcome values and the estimated values along the fitted regression line are presented in the provided graph as the vertical line segments.

Although this minimization problem can be solved by hand calculations, it becomes difficult for more than one input variable. Mathematically, the problem involves calculating the inverse of a matrix. However, other methods such as QR decomposition are used to minimize numerical round-off errors. Depending on the implementation, the storage needed to perform the OLS calculations may grow quadratically as the number of input variables grows. For many observations and many variables, the storage and RAM requirements should be carefully considered.

Note the provided equation of the fitted line. The use of the carat over  $y$ , read  $y\text{-hat}$ , is used to denote the estimated outcome for a given set of input. This notation helps to distinguish the observed  $y$  values from the fitted  $y$  values. In this example, the estimated coefficients are  $b_0 = 3.21$  and  $b_1 = 2.19$ .

## Interpreting estimated coefficients, $b_j$

### Interpreting estimated coefficients, $b_j$

$$\hat{y} = 4.0 + 2.2\text{employees} + 0.5/\text{finance} - 1.9/\text{eng} + 0.6/\text{sales}$$

- Coefficients for numeric input variables:
  - Change in outcome due to a unit change in input variable.<sup>4</sup>
  - Example:  $b_1 = 2.2$ 
    - Extra 2.2 hours per week in meetings for each additional employee managed<sup>5</sup>
- Coefficients for binary input variables:
  - Represent the additive difference from the reference level.<sup>6</sup>
  - Example:  $b_2 = 0.5$ 
    - Finance managers meet 0.5 hours per week more than engineering managers do.<sup>7</sup>
- Statistical significance of each coefficient:
  - Are the coefficients significantly different from zero?
  - For small p-values—say, less than 0.05—the coefficient is statistically significant.

<sup>4</sup>With all other input values, except the same.

For numeric variables, the estimated coefficients are interpreted in the same way as the concept of slope was introduced in algebra. For a unit change in a numeric variable, the outcome changes by the amount and in the direction of the corresponding coefficient.

A fitted linear regression model is provided for the example where the hours per week spent in meeting by managers are modeled as a function of the number of employees and the manager's functional organization. In this case, the coefficient of 2.2, corresponding to the *employees* variable, is interpreted to mean that the expected amount of time spent in meetings will increase by 2.2 hours per week for each additional employee reporting to a manager.

The interpretation of a binary variable coefficient is slightly different. When a binary variable only assumes a value of 0 or 1, the coefficient is the additive difference or shift in the outcome from the reference level. In this example, engineering is the reference level for the functional organizations. So, a manufacturing manager would be expected to spend 1.9 hours per week less in meetings than an engineering manager when the number of employees is the same.

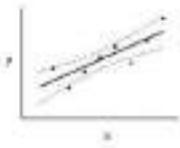
When used to fit linear regression models, many statistical software packages provide a p-value with each coefficient estimate. This p-value can be used to determine if the coefficient is significantly different than zero. In other words, the software performs a hypothesis test where the null hypothesis is that the coefficient

equals zero and the alternate hypothesis is that the coefficient does not equal zero. For small p-values—say, less than 0.05—then the null hypothesis would be rejected and the corresponding variable should remain in the linear regression model. If a larger p-value is observed, then the null hypothesis would not be rejected and the corresponding variable should be considered for removal from the model.

## Confidence and prediction intervals

### Confidence and prediction intervals

- Interval estimates provide a measure of the uncertainty in a point estimate.
- For a given value of X:
  - Confidence intervals are calculated for the mean value.
  - Prediction intervals are calculated for an individual response.
- Confidence intervals are also calculated for the coefficients:
  - If an interval straddles zero, the corresponding variable is likely not important to the model.



Bands around the line represent the confidence interval (for general stats).

DATA SOURCE: DELL INC.

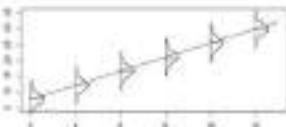
Confidence interval is the interval within which you expect the value of Y should be with a certain probability. For example, a 95 percent confidence interval would mean that you can say with 95 percent confidence that a value of Y will lie within this interval for a given value of x.

## Diagnostics—examining residuals

### Diagnostics—examining residuals

**Residual:**  
Differences between the observed and estimated outcomes.  
The observed values of the error term,  $\epsilon_i$ , in the regression model  
Expressed as:  $\epsilon_i = y_i - \hat{y}_i \quad \text{for } i = 1, 2, \dots, n$

Residuals are assumed to be normally distributed with:  
A mean of zero.  
Constant variance.

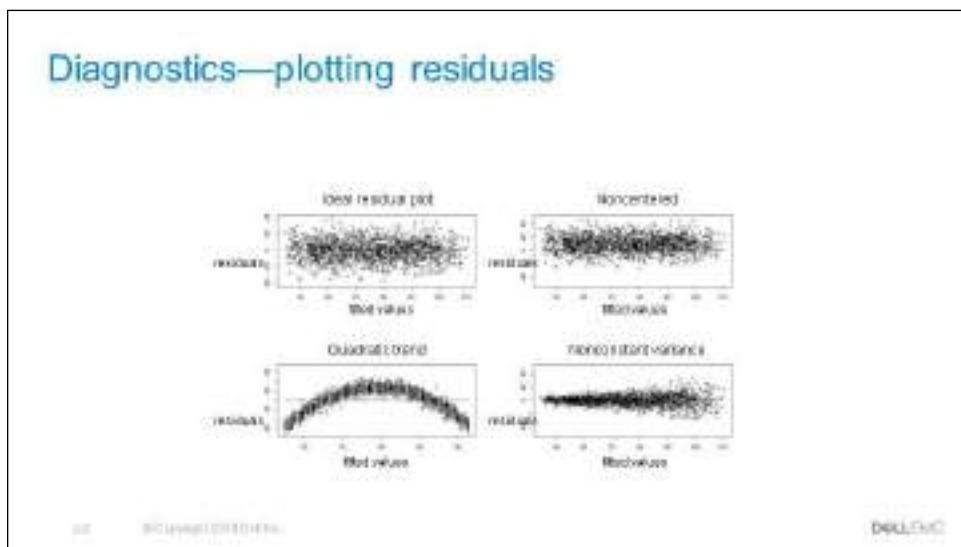


DATA SOURCE: DELL INC.

Residuals are the differences between the observed and the estimated outcomes. The residuals are the observed values of the error term in the linear regression model. In linear regression modeling, these error terms are assumed to be normally distributed with a mean of zero and a constant variance, regardless of the input variable values.

Although this normality assumption is not needed to fit a line using OLS, this assumption is the basis for many of the hypothesis tests and confidence interval calculations that statistical software packages such as R perform. The next few slides address the use of residual plots to evaluate the adherence to this assumption and to assess the appropriateness of a linear model to a given dataset.

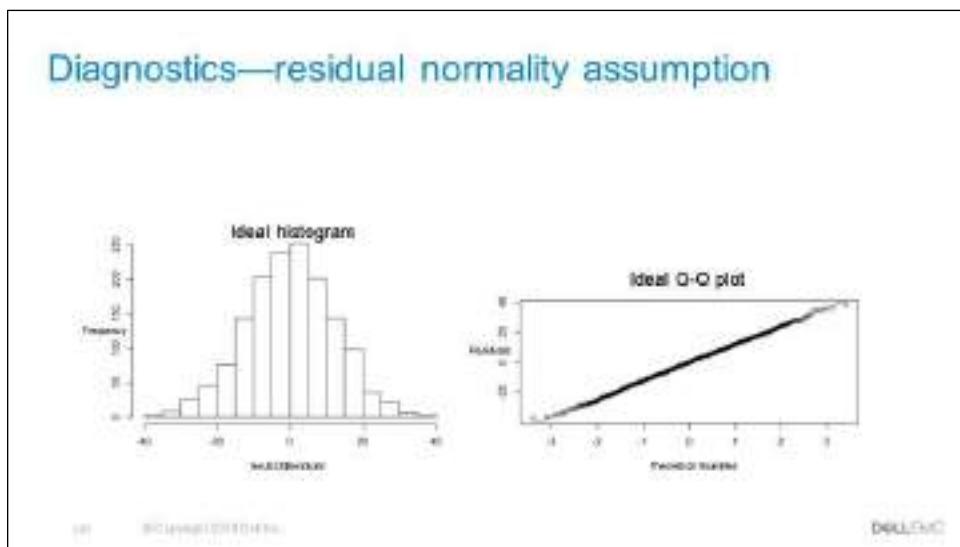
## Diagnostics—plotting residuals



When you plot the residuals against the estimated or fitted outcome values, you find that the ideal residual plot shows residuals symmetrically centered around zero with a constant variance and with no apparent trends. If the ideal residual plot is not observed, it is often necessary to add more variables to the model or transform some of the existing input and outcome variables. Common transformations include the square root and logarithmic functions.

Residual plots are also useful for identifying outliers that may require further investigation or special handling.

## Diagnostics—residual normality assumption



The provided histogram shows that the residuals are centered around zero and appear to be symmetric about zero in a bell-shaped curve, as one would expect for a normally distributed random variable. Another option is to examine a Q-Q plot that compares the observed data against the quantiles (Q) of the assumed distribution. In this example, the observed residuals follow a theoretical normal distribution represented by the line. If any significant departures of the plotted points from the line are observed, transformations, such as logarithms, may be needed to satisfy the normality assumption.

## Diagnostics—using hold-out data

### Diagnostics—using hold-out data

- Hold-out data
  - Training and testing datasets
  - Does the model predict well on data it has not seen?
- N-fold cross validation
  - Partition the data into N groups.
  - Holding out each group:
    - Fit the model
    - Calculate the residuals on the group
  - Estimated prediction error is the average over all the residuals.

© 2018 Dell Inc. All rights reserved.

DELL.COM

As discussed in Apriori diagnostics, creating a hold-out dataset before you fit the model and using that dataset to estimate prediction error is by far the easiest thing to do.

N-fold cross validation—it tells you if your set of variables is reasonable. This method is used when you do not have enough data to create a hold-out dataset. N-fold cross validation is performed by randomly splitting the dataset into N non-overlapping subsets or groups and then fitting a model using N-1 groups and predicting its performance using the group that was held out. This process is repeated a total of N times, by holding out each group. After completing the N model fits, you estimate the mean performance of the model—maybe also the variance or standard deviation of the performance.

"Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions", by Seni and Elder, provides a succinct description of N-fold cross-validation.

## Diagnostics—other considerations

### Diagnostics—other considerations

- $R^2$ 
  - The fraction of the variability in the outcome variable explained by the fitted regression model
  - Attains values from 0, indicating poorest fit, to 1, indicating perfect fit
- Identify correlated input variables
  - Pair-wise scatterplots
  - Sanity check the coefficients
    - Are the magnitudes excessively large?
    - Do the signs make sense?

141

DATA SCIENCE

DELL INC.

$R^2$ —goodness of fit metric—is reported by all standard packages. It is the fraction of the variability in the outcome variable that the fitted model explains. The definition of  $R^2$  is  $[1 - SS_{err}/S_{stot}]$ , where  $SS_{err} = \text{Sum}[(y - y_{\text{pred}})^2]$  and  $S_{stot} = \text{Sum}[(y - y_{\text{mean}})^2]$ . For a good fit, you want an  $R^2$  value near 1.

An  $R^2$  very close to 1 may suggest a trivial data set or a problem with the training data set.  $R^2$  is often used to compare two different models with the same number of input variables. The model with the better  $R^2$  score would be considered generally a better fit. When comparing two models with a different number of input variables, use pseudo- $R^2$  that contains an adjustment for the number of variables. Pseudo- $R^2$  can range from 0 to some amount greater than 1 because of adjustment for the number of variables.

Regression modeling works best if the input variables are independent of each other. A simple way to look for the correlated variables is to examine pair-wise scatterplots such as the one generated in the *Basic data analytics methods Using R* module for the Iris dataset. If two input variables,  $x_1$  and  $x_2$ , are linearly related to the outcome variable  $y$ , but are also correlated to each other, it may be only necessary to include one of these variables in the model.

After fitting a regression model, it is useful to examine the magnitude and signs of the coefficients. Coefficients with large magnitudes or intuitively incorrect signs are also indications on correlated input variables. If the correlated variables remain in

## Lesson: Linear regression

the fitted model, the predictive power of the regression model may not suffer, but its explanatory power will be diminished when the magnitude and signs of the coefficients do not make sense.

If correlated input variables must remain in the model, restrictions on the magnitudes of the estimated coefficients can be accomplished with alternative regression techniques. Ridge regression, which applies a penalty based on the size of the coefficients, is one technique that can be applied. In fitting a linear regression model, the objective is to find the values of the coefficients that minimize the sum of the residuals squared. In ridge regression, a penalty term proportional to the sum of the squares of the coefficients is added to the sum of the residuals squared. A related technique is lasso regression, in which the penalty is proportional to the sum of the absolute values of the coefficients. Both of these techniques are outside of the scope of this course.

## Linear regression—reasons to choose (+) and cautions (-)

| Linear regression—reasons to choose (+) and cautions (-)   |  |
|--|--|
| Reasons to choose (+)  | Cautions (-)   |
| Concise representation—the coefficients  | Does not handle missing values well  |
| <ul style="list-style-type: none"> <li>+ Robust to redundant or correlated variables           <ul style="list-style-type: none"> <li>– Lose some explanatory value</li> </ul> </li> <li>+ Explanatory value           <ul style="list-style-type: none"> <li>– Relative impact of each variable on the outcome</li> </ul> </li> </ul> | <ul style="list-style-type: none"> <li>+ Assumes that each variable affects the outcome linearly and additively           <ul style="list-style-type: none"> <li>– Variable transformations and modeling variable interactions can alleviate this issue</li> <li>– It is a good idea to take the log of monetary amounts or any variable with a wide dynamic range</li> </ul> </li> <li>+ Does not easily handle variables that affect the outcome in a discontinuous way           <ul style="list-style-type: none"> <li>– Step functions</li> </ul> </li> </ul> |
| Easy to score data   | <ul style="list-style-type: none"> <li>+ Does not work well with categorical attributes with many distinct values           <ul style="list-style-type: none"> <li>– For example, ZIP code</li> </ul> </li> </ul>  |

144

BIG DATA WITH PYTHON

DATA SCIENCE

The estimated coefficients provide a concise representation of the outcome variable as a function of the input variables. The estimated coefficients provide the explanatory value of the model and are used to easily determine how the individual input variables affect the outcome. Linear regression is robust to redundant or correlated variables. Although the predictive power may not be impacted, the model does lose some explanatory value in the case of correlated variables. With the fitted model, it is also easy to score a given set of input values.

A caution is that linear regression does not handle missing values well. Another caution is that linear regression assumes that each variable affects the outcome linearly and additively. If some variables affect the outcome nonlinearly and the relationships are not additive, the model often does not explain the data well. Variable transformations and modeling variable interactions can address this issue to some extent.

Hypothesis testing and confidence intervals depend on the normality assumption of the error term. To satisfy the normality assumption, a common practice is to take the log of an outcome variable with a skewed distribution for a given set of input values. Also, linear regression models are not ideal for handling variables that affect the outcome in a discontinuous way. In the case of a categorical attribute with many distinct values, the model becomes complex and computationally inefficient.

## Check your knowledge

### Check your knowledge

1. Detail the challenges with categorical values in linear regression model.
2. Describe N-Fold cross validation method used for diagnosing a fitted model.
3. List two use cases of linear regression models.
4. List and discuss two standard checks that you perform on the coefficients derived from a linear regression model.



340 800 pages of content

Dell EMC

## Check your knowledge



### Discussion

## Question / Discussion Topic:

1. Detail the challenges with categorical values in linear regression model.
2. Describe N-Fold cross validation method used for diagnosing a fitted model.
3. List two use cases of linear regression models.
4. List and discuss two standard checks that you perform on the coefficients derived from a linear regression model.

## Discussion Notes:

## Linear regression—summary

### Linear regression—summary

During this lesson, the following topics were covered:

- General description of regression models
- Technical description of a linear regression model
- Common use cases for the linear regression model
- Interpretation and scoring with the linear regression model
- Diagnostics for validating the linear regression model
- The reasons to choose (+) and cautions (-) of the linear regression model



This lesson covered these topics. Take a moment to review them.

## Lesson: Logistic regression

Introduction

Lesson: Logistic regression

DELL EMC

## Logistic regression topics

### Logistic regression topics

During this lesson, the following topics are covered:

- Technical description of a logistic regression model
- Common use cases for the logistic regression model
- Interpretation
- Diagnostics for validating the logistic regression model
- Reasons to choose (+) and cautions (-) of the logistic regression model

141

DATA SCIENCE

DELL INC.

The topics covered in this lesson are listed.

## Logistic regression

### Logistic regression

- Used to estimate the probability that an event will occur as a function of other variables
  - Example: Estimate the probability that a borrower will default as a function of credit score, income, loan amount, and any existing debt
- Input:
  - Predictor variables can be continuous or categorical
  - Outcome variable is categorical—for example, no\_default or default
- Output:
  - A set of coefficients that indicate the relative impact of each predictor variable
  - Probability that an outcome will occur
- Can be used as a classifier
  - Assign the class label based on the estimated probability
  - For example, for a high probability, assign default label

140

DATA SCIENCE

DELL INC.

Use logistic regression to estimate the probability that an event will occur as a function of other variables. For example, based on a borrower's credit score, income, loan size, and existing debts, you can estimate the probability that the borrower will default on a loan.

In logistic regression, predictor variables can be continuous or categorical. The output includes a set of coefficients that indicate the relative impact of each of the input variables. For a given set of input values, these coefficients can be used to estimate the probability of the outcome variable occurring.

Logistic regression can also be considered a classifier. Classifiers are methods to assign class labels—for example, default or no default—based on the estimated probability.

## Logistic regression use cases

### Logistic regression use cases

- Binary class outcomes:
  - A customer will purchase or not
  - A borrower will default or not
  - An applicant will accept a new job or not
  - A politician will vote yes or no
- Binary logistic regression is covered in this lesson
- Multiclass outcomes:
  - A politician will vote yes, vote no, or not vote
  - A customer will purchase, not purchase, or purchase later
- Multinomial Logistic Regression
  - Used when the dependent variable has more than two categories
  - Covered in the Advanced Methods in DSBDA course

141

DATA SCIENCE

DELL INC.

Logistic regression is especially useful if you are interested in the probability of an event, not just predicting the class labels.

This lesson covers logistic regression for binary outcomes. For multiclass outcomes, multinomial logistic regression is covered in the Advanced Methods in Data Science and Big Data Analytics course.

## Logistic regression—technical description

### Logistic regression—technical description

Based on the logistic function

$$f(y) = \frac{e^y}{1 + e^y} \quad \text{for } -\infty < y < \infty$$

where  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{p-1} x_{p-1}$

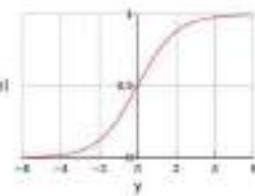
Estimates the probability of an event occurring

Consider it a linear function of predictor variables ( $y$ )

As  $y \rightarrow -\infty, f(y) \rightarrow 0$

As  $y \rightarrow +\infty, f(y) \rightarrow 1$

The logistic function is useful to estimate the probability of an outcome based on the input variables.



## Logistic regression model—typical analysis steps

**Logistic regression model—typical analysis steps**

default =  $f$  (creditScore, income, loanAmt, existingDebtFlag)

Training data: outcome variable is 0 or 1

1. If borrower does DEFAULT
0. If borrower does NOT DEFAULT

The fitted model estimates the coefficients  $\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$

Apply the fitted model to the test dataset

- if probability > 0.5, then predict the borrower will DEFAULT
- otherwise, predict the borrower will NOT DEFAULT

Compare predicted outcome to actual outcome

A different probability threshold than 0.5 may be chosen

DATA SOURCE: DELL INC.

From the available dataset, a random subset of the records is selected to build the training dataset. The remaining records are used for testing the fitted model. It is common to assign the predicted outcomes based on a probability threshold of 0.5. In other words, assign the outcome that corresponds to the probability of that particular outcome occurring. However, if detecting credit card fraud, for example, a lower threshold value, say 0.3, could be selected to ensure that the most fraudulent transactions are flagged. This approach will likely cause more nonfraudulent transactions to be incorrectly flagged as fraudulent.

## Logistic regression—visualizing model

Logistic regression—visualizing model

- Overall percentage of default: ~20%.
- Logistic regression returns a score that estimates the probability that a borrower will default.
- The graph compares the distribution of defaulters and nondefaulters as a function of the model's predicted probability, for borrowers scoring higher than 0.1.

Count

Score

Blue = Nondefaulters

Red = Defaulters

DELL INC.

Here is an example of how one might visualize the model. Logistic regression returns a score, the estimated probability, that a borrower will default. The graph compares the distribution of defaulters and nondefaulters as a function of the model's predicted probability for borrowers scoring higher than 0.1 and less than 0.98.

The graph is overlaid. Think of the blue graph—defaulters—as being transparent and in front of the red graph, representing non defaulters.

The takeaway from the graph is that the higher a borrower scores, the more likely the borrower will default.

The graph only considers borrowers who score more than 0.1 and less than 0.98 because this graph had large spikes near 0 and 1, so the graph becomes hard to read. The graph shows, however, that a fraction of low-scoring borrowers do actually default—as is seen in the overlap.

## Diagnostics—confusion matrix

**Diagnostics—confusion matrix**

|                    |                    | Actual Positive |                | Actual Negative |                |
|--------------------|--------------------|-----------------|----------------|-----------------|----------------|
|                    |                    | True Positive   | False Positive | False Negative  | True Negative  |
| Predicted Positive | Total Positive     |                 |                |                 |                |
|                    | Predicted Negative |                 |                |                 | Total Negative |
|                    | Total Positive     |                 |                |                 | Total Negative |

- Confusion matrix provides the counts of the correctly classified and incorrectly classified observations.
  - Ideally, most of the counts are in green boxes.
- In the loan default example:
  - True positive (TP):** Predicted loan will default, and the loan defaulted.
  - True negative (TN):** Predicted the loan will not default, and the loan did not.
  - False positive (FP):** Predicted the loan will default, but the loan did not.
  - False negative (FN):** Predicted the loan will not default, but the loan defaulted.
- Key metrics:
  - True Positive Rate (TPR) =  $TP / \text{Total Positive}$
  - False Positive Rate (FPR) =  $FP / \text{Total Negative}$

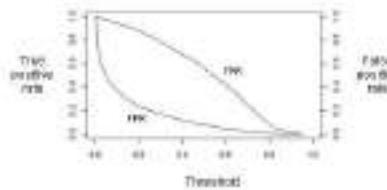
100 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800 3000 3200 3400 3600 3800 4000 4200 4400 4600 4800 5000 5200 5400 5600 5800 6000 6200 6400 6600 6800 7000 7200 7400 7600 7800 8000 8200 8400 8600 8800 9000 9200 9400 9600 9800 10000

Confusion matrix is a table that is often used to **describe the performance of a classification model** on a set of test data for which the true values are known. In general, the choice of the outcome, that is considered positive or negative, is somewhat arbitrary. In logistic regression, the outcome, that would be associated with a probability of 1, is considered positive. In this example, positive indicates a borrower defaulting on a loan.

## TPR and FPR are functions of threshold value

### TPR and FPR are functions of threshold value

- If the threshold is set to 0:
  - All observations are classified as positive.
  - Then  $\text{TPR} = \text{FPR} = 1$ .
- If the threshold is set to 1:
  - All observations are classified as negative.
  - Then  $\text{TPR} = \text{FPR} = 0$ .
- Choose threshold to obtain:
  - High TPR.
  - Low FPR.



20

DATA SCIENCE

DELL INC.

In logistic regression, the predicted outcomes depend on a chosen probability threshold value, which may be any number between 0 and 1.

- At a threshold of 0, **all** observations are classified as **positive**. Thus, all positive responses are correctly classified as positive and all negative responses are incorrectly classified as positive –  $\text{TPR} = 1$  and  $\text{FPR} = 1$ .
- At a threshold of 1, **all** observations are classified as **negative**. Thus, all positive responses are incorrectly classified as negative and all negative responses are correctly classified as negative –  $\text{TPR} = 0$  and  $\text{FPR} = 0$ .
- As the threshold increases gradually from 0 to 1 on a useful logistic model, the FPR should decrease quickly, and the TPR should remain relative close to 1. The receiver operating characteristic (ROC) curve expresses this relationship between the FPR and TPR.

## Receiver operating characteristic (ROC) curve

### Receiver operating characteristic (ROC) curve

- A perfect model has  $FPR = 0$  and  $TPR = 1$ .
- A good model has a high TPR with a low FPR.
  - Thus, the area under the curve (AUC) should be close to 1.

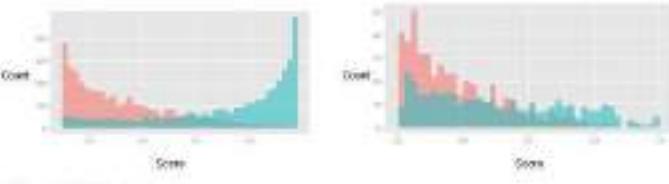
The figure shows a graph of the ROC curve. The vertical axis is labeled 'True positive rate' and ranges from 0 to 1. The horizontal axis is labeled 'False positive rate' and ranges from 0.0 to 0.6. A diagonal line from (0,0) to (1,1) represents a random classifier. A curved line above this diagonal represents a good model. The area under this curve is shaded and labeled 'Area under the curve (AUC)'.

The receiver operating characteristic (ROC) curve illustrates the relationship between TPR and FPR. As the FPR increases slightly from 0, the TPR begins to increase. In a good model, the TPR increases rapidly. Thus, the area under the curve (AUC) is used as a diagnostic measure to indicate the usefulness of the model. The closer AUC is to 1, the better the model.

## Diagnostics—plot histograms of scores

### Diagnostics—plot histograms of scores

- Logistic regression scoring gives the probability for a yes/no, given the characteristics of the predictor variables.
- To segment the prediction into yes/no, you must set a threshold. Using this threshold, you can set the population above the threshold as yes and the remaining as no.
- To set this threshold, you can decide the probability above which 90 percent of the true positives fall into that segment. The percent captured can be a business decision or through further exploring the segment selected.



The figure consists of two side-by-side histograms. Both histograms have 'Score' on the x-axis and 'Count' on the y-axis. The left histogram, labeled 'Good', shows two distinct peaks: a very tall red peak on the left and a much shorter teal peak on the right. The right histogram, labeled 'Bad', also shows two distinct peaks, but they are much closer together and of similar height, indicating significant overlap between the true and false instance score distributions.

The next diagnostic method is plotting the histogram of the scores. The graph in the left is what you saw earlier in the lesson. The graph depicts how well the model discriminates true instances from false instances. Ideally, true instances score high, and false instances score low. If so, most of the masses of the two histograms are separated. That separation is what you see in the first of these two graphs.

The graph shown at the right shows substantial overlap. The model did not predict well. This overlap means that the input variables are not strong predictors of the output.

## Diagnostic—sanity check coefficients

### Diagnostic—sanity check coefficients

- Do the signs make sense?
  - Wrong sign is an indication of correlated predictor variables.
  - Run these signs by some experts.
- Are the coefficients excessively large?
  - If so, this may indicate correlated inputs predictor variables.
  - The outcomes for a subset of the population are perfectly predicted.



SUGGESTED TOPICS

DATA VIZ

Logistic regression is an explanatory model, and the coefficients provide the necessary details.

First, check the sign of the coefficients. Do the signs make sense? For example, should the probability of defaulting increase as income increases? This idea may be counterintuitive, but the loan amount may be positively correlated with incomes. In other words, the higher the income, the higher the loan amount. Regression works best if all the drivers are independent. This does not, in fact, affect the predictive power, but the explanatory capability is compromised here.

Check if the magnitudes of the coefficients make sense. Relatively large coefficients are also an indication of strongly correlated inputs. In this case, consider eliminating or combining some variables. For example, instead of using loan amount and income in the model, consider using the ratio of the two values.

Fortunately, correlated variables do not necessarily affect the performance of the fitted logistic model. But, in this case, it is harder to interpret and explain the model.

## Other diagnostics

### Other diagnostics

- Does the model predict well on data it has not seen?
  - Evaluate the model against the testing data.
  - Avoid overfitting to training dataset.
- N-fold cross-validation
- Pseudo-R<sup>2</sup>: 1 – (residual deviance/null deviance)
  - Most software packages report residual deviance and null deviance.
    - Null deviance is based on only an intercept model ( $y = \beta_0$ ).
    - Residual deviance is based on the fitted model ( $y = \beta_0 + \beta_1 x_1 + \dots + \beta_{(j-1)} x_{(j-1)}$ ).
  - In a good-fitting model, the residual deviance should be small.
  - This is interpreted the same way R<sup>2</sup> is used in linear regression.
    - Pseudo-R<sup>2</sup> is near 1, for a good-fitting model.

100 / 300 pages completed

DELL.COM

This is all similar to linear regression. You use the hold-out data method and N-fold cross validation on the fitted model. The model should explain more than just a simple guess. Pseudo-R<sup>2</sup> is the term used in logistic regression, which is used the same way you use R<sup>2</sup> in linear regression. It is basically the fraction of the explained variability.

## Logistic regression—reasons to choose (+) and cautions (-)

| Logistic regression—reasons to choose (+) and cautions (-)                            |   |
|---|---|
| Reasons to choose (+)   | Cautions (-)  |
| Provides relative impact of each predictor variable on the likelihood of the outcome. | Assumes that a linear combination of the predictor variables helps to estimate the probability of the outcome.                  |
| Robust with correlated variables.   | Correlated variables reduce the explanatory value of the model.   |
| Concise representation with the coefficients.   | Cannot handle variables that affect the outcome in a discontinuous way.   |
| Returns probability estimates of an event.  | Sigmoid functions.<br>Does not work well with categorical predictor variables with many distinct values—for example, zip codes. |

Based on the signs and magnitudes of the coefficients, the effect a change in the values of the predictor variables on the likelihood of the outcome can be explained. Logistic regression handles with correlated variables. In this case, the prediction is not impacted, but some explanatory value is lost with the fitted model. Logistic regression provides the concise representation of the outcome with the coefficients, and it is easy to score new data with this model. Logistic regression returns probability estimates of an event.

Cautions (-) are that the logistic regression still assumes that the probability of the outcome is proportional to a linear combination of the predictor variables. Also, when you have a categorical predictor variable with many distinct values, the model becomes complex and computationally inefficient.

## Check your knowledge

**Check your knowledge**

1. What is meant by a binary outcome?
2. How is ROC curve used to diagnose the effectiveness of the logistic regression model?
3. What is Pseudo-R<sup>2</sup> and what does it measure in a logistic regression model?
4. Compare and contrast linear and logistic regression methods.



© 2018 Dell Inc. All rights reserved.

## Check your knowledge



### Discussion

### Question / Discussion Topic:

1. What is meant by a binary outcome?
2. How is ROC curve used to diagnose the effectiveness of the logistic regression model?
3. What is Pseudo-R<sup>2</sup> and what does it measure in a logistic regression model?
4. Compare and contrast linear and logistic regression methods.

### Discussion Notes:

## Logistic regression—summary

### Logistic regression—summary

During this lesson, the following topics were covered:

- Technical description of a logistic regression model
- Common use cases for the logistic regression model
- Interpretation and scoring with the logistic regression model
- Diagnostics for validating the logistic regression model
- Reasons to choose (+) and cautions (-) of the logistic regression model



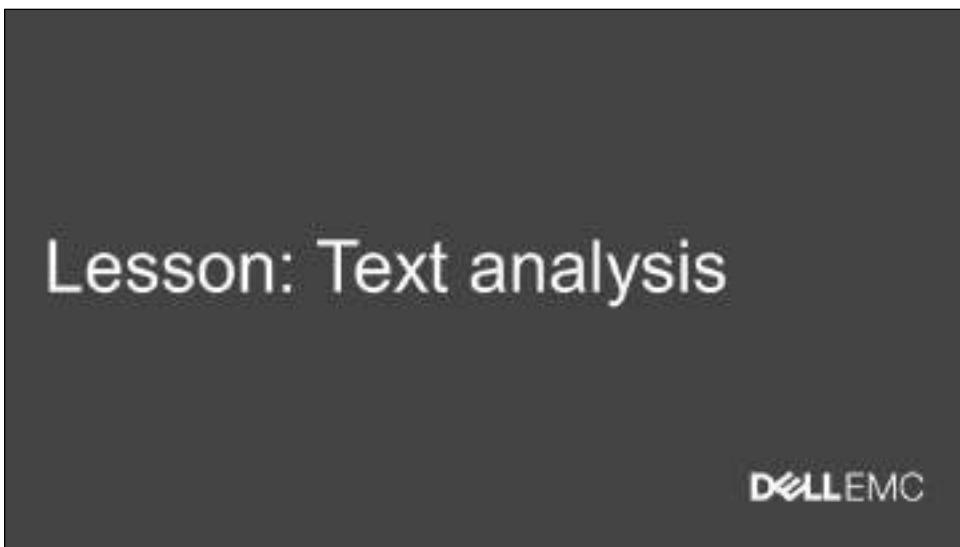
DATA SCIENCE

DELL EMC

This lesson covered these topics. Take a moment to review them.

## Lesson: Text analysis

### Introduction



## Text analysis

### Text analysis

During this lesson, the following topics are covered:

- Challenges with text analysis
- Key tasks in text analysis
- Definition of terms used in text analysis
  - Term frequency, inverse document frequency
- Representation and features of documents and corpus
- Use of regular expressions in parsing text
- Metrics used to measure the quality of search results
  - Relevance with TF-IDF

300

DATA SCIENCE

DELL INC.

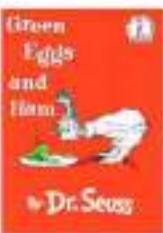
The topics covered in this lesson are listed.

## Text analysis, cont.

### Text analysis, cont.

Encompasses the processing and representation of text for analysis and learning tasks.

- Analyzing tweets
  - Twitter currently generates 500MM tweets per day
  - It generates tweets in close to 40 languages
  - Distinct number of words can be close to 100,000
- High-dimensionality
  - Every distinct term is a dimension
  - Green Eggs and Ham: A 50-D problem!
- Data is unstructured

DATA SOURCE: DELL INC.

Text analysis is essentially the processing and representation of data that is in text form to analyze it and learn new models from it.

The main challenge in text analysis is the problem of high dimensionality. When you analyze a document, you find that every possible word in the document represents a dimension.

Consider the book 'Green Eggs and Ham' by Dr. Seuss, which he wrote in response to a challenge to write a book with just 50 different words.  
(Reference: [https://en.wikipedia.org/wiki/Green\\_Eggs\\_and\\_Ham](https://en.wikipedia.org/wiki/Green_Eggs_and_Ham))

Even this book represents a 50-dimension problem, if you consider vectors in a text space.

The other major challenge with text analysis is that the data is unstructured.

## Text analysis—problem-solving tasks

### Text analysis—problem-solving tasks

- Parsing
  - Impose a structure on the unstructured or semi-structured text
  - Useful for later analysis
- Search and retrieval
  - Which documents have this word or phrase?
  - Which documents are about this topic or this entity?
- Text mining
  - “Understand” the content
  - Clustering and classification
- Tasks are not an ordered list:
  - Does not represent process
  - Perform tasks based on the problem you want to address

CC BY-SA 4.0 Dell EMC

Dell EMC

The problem-solving tasks in text analysis include three important steps—namely, parsing, search and retrieval, and text mining.

**Parsing** is the process step that takes the un-structured or a semi-structured document and imposes a structure for the downstream analysis. Parsing is basically reading the text, which could be weblog, an RSS feed, an XML or HTML file, or a word document. Parsing decomposes what is read in and renders it in a structure for the subsequent steps.

After parsing is completed, the problem focuses on **search and retrieval** of specific words or phrases or in finding a specific topic or entity—a person or a corporation—in a document or a corpus, or body of knowledge. All text representation takes place implicitly in the context of the corpus. All search and retrieval is something commonly performed with search engines such as Google. Most of the techniques used in search and retrieval originated from the field of library science.

With the completion of these two steps, the output generated is a structured set of tokens or a bunch of key words that were searched, retrieved, and organized. The third task is **mining the text** or understanding the content itself. In this step, instead of treating the text as set of tokens or keywords, you can derive meaningful insights into the data pertaining to the domain of knowledge, the business process, or the problem you are trying to solve.

## Lesson: Text analysis

Many of the techniques mentioned in the previous lessons, such as clustering and classification, can be adapted to the text mining, with the proper representation of the text. You could use K-means clustering or other methods to tie the text into meaningful groups of subjects. Sentiment Analysis and Spam filtering are examples of classification tasks in text mining—recall Spam filtering was listed as a prominent use case for Naïve Bayesian Classifier. In addition to traditional statistical methods, Natural Language processing methods are also used in this phase.

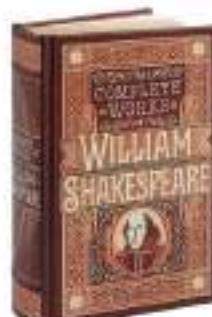
It should be noted the list of tasks are not ordered. One generally starts with the parsing, either with the intention of compiling them into a searchable corpus or catalog—maybe after some analytical tasks such as tagging or categorization—OR specifically for text mining. So, it is not a process. It is a set of things that go into the text analysis task, or maybe a tree, where you start with parsing, and go down to either search or to text mining.

The rest of this lesson details each of these steps.

## Example—term frequency

### Example—term frequency

- Term frequency—Count of words or terms in each document or across documents
- Calculate the term frequency of the words in the complete works of Shakespeare
  - Total words—0.88 MB, Distinct words—18,000
- Each distinct work is stored as a document. Each document contains words.
- Within the works, there is a possibility that some of documents are dramatis personae—a document with character details.
- Find the short documents that are dramatis personae and filter them from the dataset.



DATA SCIENCE

Here, you see an example of finding the term frequency in Shakespeare's complete works.

Shakespeare has completed over 150 plays. Some of them are well known, such as Romeo and Juliet, Hamlet, and Macbeth.

Here, try to analyze how many times the words are repeated across plays.

When getting an electronic version of the works into the RStudio environment, first analyze the correctness of the works. In doing so, you might see copyright information, headers, and footers. You must ensure you eliminate such information before proceeding with the exercise.

Also, in this version, you see dramatis personae as separate documents. Since this only lists the characters in the play, you must delete them to get an accurate representation of term frequency.

## Representing corpus—collection of documents and features

### Representing corpus—collection of documents and features

- A corpus is a collection of text documents.
  - Group of news articles
  - A set of emails or tweets
- Corpus metrics
  - Volume
  - Corpus-wide term frequencies
  - Inverse document frequency (IDF)
- Challenge: a corpus is often dynamic
  - Indexes and metrics must be updated continuously

DATA SCIENCE

DELL INC.

It is important that you not only create a representation of the document but that you also represent a corpus. What is the representation of a corpus?

Now that you have collected the reviews and turned them into the proper representation, archive them in a searchable archive for future reference and research.

The other corpus metrics such as volume and corpus-wide term frequency, which specifies how the terms are distributed across the corpus, help with the downstream analysis of classification and searching. Search algorithms also use inverse document frequency, defined later in this lesson.

A fact that many people do not think about is that documents are often only relevant in the context of a corpus, or a specific collection of documents. Sometimes, this fact is obvious, as in the case of search or retrieval. It is less obvious in the case of classification—for example, spam filtering, sentiment analysis. But, even in that case, the classifier has been trained on a specific set of documents. And, the underlying assumption of all classifiers is that they will be deployed on a population that is similar to the population that they were trained on.

A primary challenge in text analysis and search is that a corpus changes constantly over time. Not only do new documents get added—which means the metrics and indexes must be updated—but word distributions can change over time, which will

reduce the effectiveness of classifiers and filters, if they are not retrained. Think about spam filters.

The corpus representation presented here is primarily oriented towards search and retrieval, but some of the metrics, such as IDF, can also be relevant to classification.

## Text classification—parsing and tokenizing

### Text classification—parsing and tokenizing

- Tokenization is the task of separating words from the body of text.
- Tokens are typically words, but can be fragments of words
- Not as straightforward as it seems
  - Corpus can help you by removing punctuations, numbers, and stop words so that they are not tokenized.
  - If space is used for tokenizing, "day." could be one form of word.
  - If punctuation is used for tokenizing, you get "day" and ". ". But, It also converts "we'll" to "we" and "ll".
  - You must find the right token.
- Each language might need a new method of tokenization.
- Case folding is another method used for tokenizing

000

DATA SCIENCE

DELL INC.

Tokenizing is process of converting a sentence into a group of words. Any kind of separators can be used to find all the words in the sentence.

If space is used as the separator of the tokens, then "day." and "day" would be considered different. One way to fix it is to remove the period at the end of all sentences. Also, you can tokenize on punctuations.

Tokenizing based on punctuation marks might not be well suited to certain scenarios. For example, if the text contains contractions such as "we'll", tokenizing based on punctuation splits them into separated words "we" and "ll". For words such as "can't", the output would be "can" and "t".

Tokenization is a much more difficult task than one may expect. For example, should words such as state-of-the-art, Wi-Fi, and San Francisco be considered one token or more?

Another text normalization technique is called *case folding*, which reduces all letters to lowercase—or the opposite, if applicable.

If implemented incorrectly, case folding may reduce or change the meaning of the text and create more noise—for example, when "General Motors" becomes "general" and "motors".

## Extract and represent text

**Extract and represent text**

Document to handle:  
A structure for analysis.

**"Bag of words"**  
Common representation:  
A vector with one dimension for every unique term in a text.

**Term frequency (tf)**  
NUMBER OF DOCUMENTS  
Geometric mean search, its utilization.

Representation density:  
Term table—full, dense  
Vector model “tf”, “tf”  
One-dimensional column  
Training  
“From” = “Phone”

We know what we are, but not what we may be.

Consider this situation in vector form:

|        |   |
|--------|---|
| love   | 3 |
| Hamlet | 2 |
| the    | 0 |
| are    | 1 |
| we     | 0 |
| know   | 1 |
| what   | 0 |
| not    | 0 |
| what   | 0 |
| we     | 0 |
| may    | 0 |
| be     | 0 |

DELL INC.

You are now in Step 2. You have parsed all your data feeds and collected the phrases and words, and you are ready to represent what you collected in a structured manner for downstream analysis.

The most common representation of the structure is known as the “bag of words.” The bag of words is a vector with one dimension for every unique term in the space.

Here, too, is an introduction to the term “term frequency” (tf), which is the number of times a term occurs in a vector.

Obviously, the vector is **very** high-dimensional, as you invariably end up with a significant number of unique words in a document. Bag of words is a common representation, and it is suited well for search and classification. There are more sophisticated representations for sophisticated algorithms.

In the example shown here, you parsed a quote in Hamlet, “We know what we are, but not what we may be”—you are only showing part of your vector space.

Now, count the occurrences of the words in the text parsed and the number of times the word is repeated. Then, store the word count as a part of the vector representation. In this example, you see “bPhone” mentioned once and “love” mentioned twice.

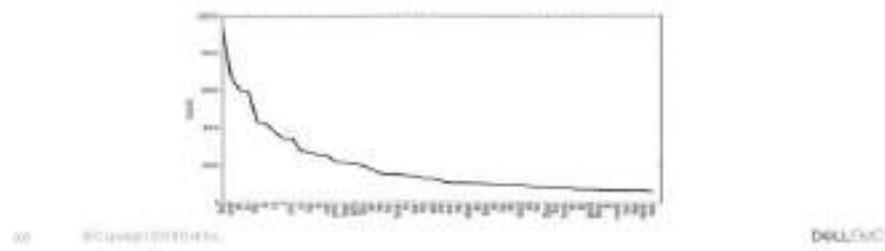
## Lesson: Text analysis

To reduce the dimensionality, do not include all the words in the English language. Normally, you ignore some “stop” words such as “the”, “a”, and so on. Other methods include stemming the words and avoiding pronouns in the term space. Vector space must be managed in a way so that it only contains words that are essential for the analysis. Stemming is completed based on the context and corpus. In an unstructured document, techniques such as parts of speech tagging are used for parsing.

## Extract and represent text, cont.

### Extract and represent text, cont.

- Using single words as identifiers with the bag-of-words representation, the term frequency of each word can be calculated.
- As an example, here is a plot for terms with highest counts from Shakespeare's Hamlet.



Using single words as identifiers with the bag-of-words representation, the *term frequency* (TF) of each word can be calculated. Term frequency represents the weight of each term in a document, and it is proportional to the number of occurrences of the term in that document.

You can generate a frequency table that shows the ranking of the words, with the most frequent word having the best rank.

You can plot the frequencies to get an understanding of the distribution of words in the play. You can practically apply the same concepts to know the frequency of words that are for and against a company or product.

## Computing relevance—term frequency

### Computing relevance—term frequency

- Along with the Bag-of-words method, you can also use this function to calculate weights.
- Assign each term in a document a weight for that term.
- The weight of a term  $t$  in a document  $d$  is a function of the number of times  $t$  is displayed in  $d$ .

– The weight can be set to the number of occurrences of  $t$  in  $d$ :

$$tf(t, d) = \text{count}(t, d)$$

...  
...

...  
...

DELL EMC

Here, you see a simple example of how relevance might be computed.

First, call up all the documents that have any of the terms from the query, and count how many times each term occurs. The more often a term is mentioned in the document, the more relevant the document is.

Obviously, there are ways to improve this method. For example, one might prefer documents that include ALL the terms, not just any. Also, one might want to limit the weight accorded to any one term—for example, "Spam spam spam spam, wonderful spam..."

## Inverse document frequency (IDF)

### Inverse document frequency (IDF)

$$\text{IDF}(t) = \log \frac{N}{df(t)+1}$$

- N: Number of documents in the corpus
- df(t): Number of documents in the corpus that contain a term t
- Measures term uniqueness in corpus
  - "love" vs. "truth"
- Indicates the importance of the term
  - Search—relevance
  - Classification—discriminatory power

100

DATA SCIENCE

DATA SCIENCE

Inverse document frequency enables you to improve your search algorithm.

IDF measures the uniqueness of a term in the corpus. If a term shows up only in 10 percent of the documents, then it is unique. If a term shows up in 90 percent of the documents, then it is not all that unique. It indicates the importance of the term—which, in this case, is displayed in 10 percent of documents—and **provides relevance to the search by weighting the rare term higher**.

In a corpus of Shakespeare plays, the word "love" is probably common. The term "truth" is probably less common, as it concerns only some of the plays. So, it is an important term when it shows up in a query—it discriminates relevant documents better than "love" does—and potentially is distributed differently in each category of plays. IDF reflects the fact that "truth" is potentially an interesting feature of a document.

## More possibilities with text analysis

### More possibilities with text analysis

- Topic tagging—segment the text documents into the specific categories
  - Segment books into the right categories. There are 300,000 titles published in 2013 in the US only.
  - In the current example, the plays can be categorized into comedies, tragedies, and histories.
- Sentiment analysis refers to tasks that use statistics and natural language processing to mine opinions to identify and extract subjective information from texts.
  - This analysis is performed on Twitter to determine overall opinion on a particular trending topic.
  - Companies and brands often utilize sentiment analysis to monitor brand reputation across social media platforms.

10

DATA SCIENCE

DELL INC.

Topic modeling provides tools to automatically organize, search, understand, and summarize from vast amounts of information. Topic models are statistical models that examine words from a set of documents, determine the themes over the text, and discover how the themes are associated or change over time.

The process of topic modeling can be simplified to the following:

1. Uncover the hidden topical patterns within a corpus.
2. Annotate documents according to these topics.
3. Use annotations to organize, search, and summarize texts.

Many companies, such as Amazon or Shopping.com, rely on teams of hand-taggers to create training corpora to jump-start efforts in automated categorization. Hand-tagged data is slow to collect, and is prone to fatigue errors and inconsistent—subjective—tagging on the part of the taggers.

In the case of sentiment analysis, one could try creating training corpora based on sites that have quantitative ratings for the products. The resulting classifiers run the risk of only being effective on the reviews from sites that they came from, or for reviews from that product category, because of idiosyncratic terminology of the website community or the product category. As an example, "lightweight" is a positive adjective for laptops, but not necessarily for wheelbarrows, or books.

Classifiers built from reviews would almost definitely not work on tweets or blog comments.

Using unsupervised methods to cluster the documents, and then assigning labels based on whether the sampled documents from a cluster are positive or negative might work. But, since the cluster is not built specifically on sentiment, it may not partition on sentiment.

There are other things you can do to track sentiment, besides classification. For instance, you can track the frequency with which certain words appear in reviews of your products, and then let a human decide if the overall trend looks positive or negative. The point of this discussion is not to cover all the possible ways of text mining, but to cover the basic concepts and issues.

## Natural language processing

### Natural language processing

- Natural language processing (NLP) is the capacity of a computer to "understand" natural language text at a level that allows for meaningful interaction between the computer and a person working in a particular application domain.
- Most companies use NLP extensively..
- **Usage of NLP**
  - Social media monitoring
  - Text analytics
  - Formulate responses using natural language
  - Sentiment classification
  - Chatbots

DATA SCIENCE & BIG DATA ANALYTICS

DELL EMC

NLP is a range of computational techniques for analyzing and representing naturally occurring speech at one or more levels of linguistic analysis, to achieve human-like language processing for a range of tasks or applications.

The goal of natural language processing is to enable computers to perform useful tasks that involve human natural languages. It is concerned with all levels of interactions between humans, or humans and computers.

Some sample tasks of natural language processing include:

- Normalizing text with tokenization and stemming before the analysis
- Extracting useful information from unstructured text with disambiguation and part-of-speech (POS) tagging
- Conducting sentiment analysis on the web to gauge positive or negative feelings related to a specific topic
- Understanding the meaning of the text including parsing and semantic analysis
- Automated scoring of student essays—in the SAT and GRE tests
- Chatbot—the computer programs you can talk to through messaging apps, chat windows, or voice-calling apps

Next, you will see an example related to understanding the meaning of texts to illustrate the ambiguity issues that NLP encounters.

## Tough nut to crack—NLP

### Tough nut to crack—NLP

- NLP must overcome ambiguity and volume of data for it to understand the sentence.
- Types of ambiguity
  - Lexical ambiguity – words having multiple meanings
  - Syntactic ambiguity – sentence having multiple parse trees
  - Semantic ambiguity – sentences having multiple meanings
  - Anaphoric ambiguity – phrase or word that is previously mentioned but having a different meaning
- A complete guide to NLP is now available in the Advanced Methods in Data Science and Big Data Analytics course

DATA SCIENCE

DELL INC.

One major difficulty is that you might not consciously understand language yourselves. The second major difficulty is ambiguity.

When you think of a linguistic concept such as a word or a sentence, those ideas seem to be simple and well-formed. But, in reality, there are many borderline cases that can be difficult to figure out.

"Ground" has tons of meanings as a verb, and even more as a noun. To understand what a sentence means, you must understand the meaning of the words, and that is no simple task.

For humans, if a language is known, understanding a sentence is effortless. When you read a web page with lists, tables, run-on sentences, newly made-up words, nouns used as verbs, and sarcasm, you "get it" immediately, usually without having to work at it.

To make a computer understand all the different ways humans use words, you must build an engine that understands the intricacies in the usage of different words. NLP is the platform for building these engines. A complete guide to NLP is now available in the Advanced Methods in Data Science and Big Data Analytics course.

## Challenges—text analysis

### Challenges—text analysis

- Finding the right structure for your unstructured data
- Very high dimensionality
- Thinking about your problem the right way



DATA & DOCUMENTATION

DELL EMC

As a recap, there are several key challenges with text analysis.

As you saw in the module on the data analytics lifecycle, the most challenging aspect of data analytics problems often is not the statistics or mathematical algorithms; it is formulating the problem, getting the data, and preparing the data. This aspect is especially true for text analysis.

## Check your knowledge

### Check your knowledge

1. What are the two major challenges in the problem of text analysis?
2. Why is the corpus metrics dynamic. Provide an example and a scenario that explains the dynamism of the corpus metrics.
3. How does tf-idf enhance the relevance of a search result?
4. List and discuss a few methods that are deployed in text analysis to reduce the dimensions.



## Check your knowledge



### Discussion

## Question / Discussion Topic:

1. What are the two major challenges in the problem of text analysis?
2. Why is the corpus metrics dynamic. Provide an example and a scenario that explains the dynamism of the corpus metrics.
3. How does tf-idf enhance the relevance of a search result?
4. List and discuss a few methods that are deployed in text analysis to reduce the dimensions.

## Discussion Notes:

## Text analysis—summary

### Text analysis—summary

During this lesson, the following topics were covered:

- Challenges with text analysis
- Key tasks in text analysis
- Definition of terms used in text analysis
  - Term frequency, inverse document frequency
- Representation and features of documents and corpus
- Metrics used to measure the quality of search results
  - Relevance with TF-IDF



This lesson covered these topics. Take a moment to review them.

## Lesson: Naïve Bayes

Introduction



Lesson: Naïve Bayes

DELL EMC

## Naïve Bayes

### Naïve Bayes

During this lesson, the following topics are covered:

- Theoretical foundations of the Naïve Bayes classifier
- Use cases
- Evaluating the effectiveness of the classifier
- Reasons to choose (+) and cautions (-)

The topics covered in this lesson are listed.

## Classifiers

The screenshot shows a presentation slide with a light gray header bar containing the title 'Classifiers'. The main content area has a white background with a thin gray border. At the top left of this area, the word 'Classifiers' is written in a teal color. Below it is a large rectangular box with a thin gray border, containing three questions in black text: 'Where in the catalog should I place this product listing?', 'Is this email spam?', and 'Will the customer buy the product?'. To the right of these questions is a list of classifiers under two bullet points: 'Classification' and 'Commonly used classifiers'. The 'Classification' point has two sub-points: 'Assign labels to objects.' and 'Usually supervised - training dataset of preclassified observations'. The 'Commonly used classifiers' point has three sub-points: 'Naïve Bayes', 'Decision Trees', and 'Logistic Regression'. At the bottom left of the slide, there is some very small, faint text that appears to be a URL or a reference. At the bottom right, the Dell logo is visible.

- Classification
  - Assign labels to objects.
  - Usually supervised - training dataset of preclassified observations
- Commonly used classifiers
  - Naïve Bayes
  - Decision Trees
  - Logistic Regression

The primary task that classifiers perform is to assign labels to objects. Labels in classifiers are predetermined, unlike in clustering, where you discover the structure and assign labels. Classifier problems often apply supervised learning methods to a training dataset of preclassified examples. For example, retailers use classifiers to assign proper catalog entry locations for their products. Also, the identification of emails as spam is a useful application of classifier methods.

## Naïve Bayes classifier approach

### Naïve Bayes classifier approach

- **Based on the observed object attributes:**
  - Naively assumed to be conditionally independent of each other
  - Class label probabilities are determined using Bayes' Law
  - Determine the most probable class label for each object
- **Example:**
  - Classify an object based on its attributes {shape, color, weight}
  - Given an object that is {spherical, yellow, <60 grams}
    - » Tennis ball, green spherical, yellow, <60 grams) = 0.32
    - » Apple, green spherical, yellow, <60 grams) = 0.02
    - » Pinbowling ball, green spherical, yellow, >60 grams) = 0.0000001
- **Input variables are discrete, categorical**
- **Output:**
  - Probability score for each possible class label
    - » Proportional to the true probability
  - Assigned class label, based on the highest probability score

000

DATA SCIENCE

DELL EMC

**The Naïve Bayes classifier is a probabilistic classifier based on Bayes' Law and naïve conditional independence assumptions.** In simple terms, a Naïve Bayes classifier assumes that the presence—or absence—of a particular feature of a class is unrelated to the presence—or absence—of any other feature.

For example, an object can be classified into a particular category based on its attributes such as shape, color, and weight. A reasonable classification for an object that is spherical, yellow, and less than 60 grams in weight may be a tennis ball. Even if these features depend on each other or upon the existence of the other features, a Naïve Bayes classifier considers all of these properties to independently contribute to the probability that the object is a tennis ball.

The input variables are discrete—categorical—but there are variations to the algorithms that work with continuous variables, as well. For this lesson, only discrete input variables are considered. Although weight may be considered a continuous variable, in the tennis ball example, weight was grouped into intervals to treat weight as a categorical variable.

The output typically returns a probability score and class membership. The outputs from most implementations are log probability scores for the class; assign the class label that corresponds to the highest log probability score.

## Naïve Bayes—use cases



Naïve Bayes—use cases

- Insurance fraud detection
- Text classification
  - Spam filtering
  - Document classification
- Medical diagnosis
- Applicable for cases with:
  - Many input variables and values
  - Multiple class labels

Naïve Bayes is useful for classification problems involving many input variables with many possible values. For example, the choice of words used in emails could be used to classify the emails as spam or not spam. The Naïve Bayes approach is efficiently handles the thousands of possible words that may appear in the emails.

In the auto insurance industry, based on a training dataset with attributes such as:

- Driver's rating, vehicle age,
- Vehicle price,
- Is it a claim by the policy holder,
- Police report status, and
- Claim genuine flag

A new claim can be classified as genuine or not.

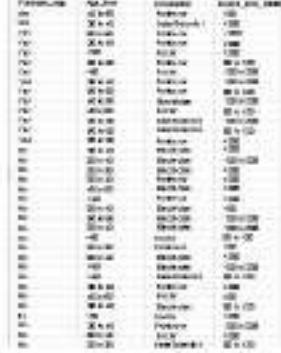
### References:

- Insurance fraud detection: [www.cisjournal.org/archive/vol2no4/vol2no4\\_1.pdf](http://www.cisjournal.org/archive/vol2no4/vol2no4_1.pdf)
- Spam filtering: [en.wikipedia.org/wiki/Bayesian\\_spam\\_filtering](https://en.wikipedia.org/wiki/Bayesian_spam_filtering)

## Build training dataset to predict customer purchase

**Build training dataset to predict customer purchase**

- Predict if the customer will purchase the product based on their profile:
  - Age bins
  - Occupation
  - Income tier
- Note: Continuous variables are transformed into categorical variables.



CustomerID  
Age Group  
Occupation  
Income Tier

Dell EMC

Here, you see a specific use case example. The training dataset consists of attributes: customer's age tier, occupation, and income tier. They are represented as categorical variables, which are well suited for naïve Bayes.

With this training set, you want to predict whether a new customer will purchase or not. This problem could be solved with logistic regression, as well. If there are multiple levels for the outcome you want to predict, then Naïve Bayesian Classifier is a better solution.

Next, you will go through the technical basis for Naïve Bayesian Classifiers and will revisit this credit dataset later.

## Conditional probability

### Conditional probability

The probability of event C occurring given event A has occurred  
Denoted as  $P(C | A)$

Example:

A fair 6-sided die is thrown

Let  $A = \{\text{an even number is rolled}\}$

If  $C = \{\text{a 3 is rolled}\}$ , then  $P(C | A) = 0$

If  $C = \{\text{a 4 is rolled}\}$ , then  $P(C | A) = 1/3$

Knowing that A occurred, provides information about the probability of C

Formal definition:

$$P(C | A) = \frac{P(A \cap C)}{P(A)} \quad \text{for } P(A) > 0$$

where  $P(A \cap C)$  denotes probability of events A and C occurring

## Derivation of Bayes' Law

### Derivation of Bayes' Law

By definition of conditional probability,

$$P(C | A) = \frac{P(A \cap C)}{P(A)} \quad (1)$$

Alternatively,

$$P(A | C) = \frac{P(A \cap C)}{P(C)} \rightarrow P(A \cap C) = P(A | C)P(C) \quad (2)$$

Substituting back into the definition yields:

$$P(C | A) = \frac{P(A | C)P(C)}{P(A)}$$

Known as Bayes' Law

A conditional probability can be expressed as a function of another conditional probability.

300 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800 3000 3200 3400 3600 3800 4000 4200 4400 4600 4800 5000 5200 5400 5600 5800 6000 6200 6400 6600 6800 7000 7200 7400 7600 7800 8000 8200 8400 8600 8800 9000 9200 9400 9600 9800 10000

DELL EMC

For  $P(A)$  and  $P(C)$  both greater than zero, the two conditional probabilities,  $P(C | A)$  and  $P(A | C)$  can be expressed.

Solving second equation for  $P(A \cap C)$  and substituting into the first equation expresses  $P(C | A)$  in terms of  $P(A | C)$ ,  $P(A)$ , and  $P(C)$ . This result is known as Bayes' Law. Often, one conditional probability is easier to calculate than the other. In such cases, it is useful to utilize Bayes' Law.

## Application of Bayes' Law

**Application of Bayes' Law**

**Scenario**

John flies frequently and likes to upgrade his seat to first class.

If John arrives at least two hours early, then he will get the upgrade 75 percent of the time.

Otherwise, he will get the upgrade 35 percent of the time.

John arrives at least two hours early only 40 percent of the time.

Suppose that John did not receive an upgrade on his most recent attempt.

**What is the probability that he arrived late?**

$$P(\text{Late} | \text{No Upgrade}) = \frac{P(\text{No Upgrade} | \text{Late})P(\text{Late})}{P(\text{No Upgrade})}$$

$$= \frac{(1 - 0.35)(1 - 0.40)}{1 - (0.40 * 0.75 + 0.60 * 0.35)} \approx 0.80$$

DATA SCIENCE

Here is an example of the application of Bayes' Law.

## Apply Naïve assumption and remove constant

### Apply Naïve assumption and remove constant

For observed attributes  $A = (a_1, a_2, \dots, a_m)$ , compute

$$P(C_i | A) = \frac{P(a_1, a_2, \dots, a_m | C_i)P(C_i)}{P(a_1, a_2, \dots, a_m)} \quad i = 1, 2, \dots, n$$

and assign the classifier  $C_i$  with the largest  $P(C_i | A)$

Two simplifications to the calculations

Apply naïve assumption - each  $a_j$  is conditionally independent of each other, then:

$$P(a_1, a_2, \dots, a_m | C_i) = P(a_1 | C_i)P(a_2 | C_i) \cdots P(a_m | C_i) = \prod_{j=1}^m P(a_j | C_i)$$

Denominator  $P(a_1, a_2, \dots, a_m)$  is a constant and can be ignored

The general approach is to assign the classifier label,  $C_i$ , to the object with attributes  $A = (a_1, a_2, \dots, a_m)$  that corresponds to the largest value of  $P(C_i | A)$ .

The probability that a set of attribute values  $A$ —comprised of  $m$  variables  $a_1$  through  $a_m$ —should be labeled with a classification  $C_i$  is equal to the probability that of the set of variables  $a_1$  through  $a_m$  given  $C_i$  is true, times the probability of  $C_i$  all divided by the probability of the set of attribute values  $a_1$  through  $a_m$ .

The conditional independence assumption is that the probability of observing the value of a particular attribute given  $C_i$  is independent of the other attributes. This Naïve assumption simplifies the calculation of  $P(a_1, a_2, \dots, a_m | C_i)$  as shown on the slide.

Since  $P(a_1, a_2, \dots, a_m)$  is displayed in the denominator of  $P(C_i | A)$ , for all values of  $i$ , removing the denominator has no impact on the relative probability scores and simplifies calculations. Next, apply these two simplifications to the calculations, to build the Naïve Bayesian Classifier.

## Building Naïve Bayesian classifier

### Building Naïve Bayesian classifier

Applying the two simplifications

$$P(C_i | a_1, a_2, \dots, a_n) \propto \left( \prod_{j=1}^n P(a_j | C_i) \right) P(C_i) \quad i = 1, 2, \dots, n$$

To build a Naïve Bayesian Classifier, collect the following statistics from the training data:

$P(C_i)$  for all the class labels

$P(a_j | C_i)$  for all possible  $a_j$  and  $C_i$

Assign the classifier label  $C_i$  that maximizes the value of

$$\left( \prod_{j=1}^n P(a_j | C_i) \right) P(C_i) \quad i = 1, 2, \dots, n$$

Applying the two simplifications,  $P(C_i | a_1, a_2, \dots, a_m)$  is proportional to the product of the various  $P(a_j | C_i)$ , for  $j=1,2,\dots,m$ , times  $P(C_i)$ . From a training dataset, these probabilities can be computed and stored for future classifier assignments. This process can be applied to the credit applicant example.

## Naïve Bayesian classifiers for product purchase example

| Purchased_Seq | Age (years) | Occupation     | Home_Lock_100 |
|---------------|-------------|----------------|---------------|
| 1             | 25          | Electrician    | 0.00          |
| 2             | 30          | Electrician    | 0.00          |
| 3             | 35          | Data Scientist | 0.00          |
| 4             | 40          | Electrician    | 0.00          |
| 5             | 45          | Data Scientist | 0.00          |
| 6             | 50          | Electrician    | 0.00          |
| 7             | 55          | Data Scientist | 0.00          |
| 8             | 60          | Electrician    | 0.00          |
| 9             | 65          | Data Scientist | 0.00          |
| 10            | 70          | Electrician    | 0.00          |
| 11            | 75          | Data Scientist | 0.00          |
| 12            | 80          | Electrician    | 0.00          |
| 13            | 85          | Data Scientist | 0.00          |
| 14            | 90          | Electrician    | 0.00          |
| 15            | 95          | Data Scientist | 0.00          |
| 16            | 100         | Electrician    | 0.00          |
| 17            | 105         | Data Scientist | 0.00          |
| 18            | 110         | Electrician    | 0.00          |
| 19            | 115         | Data Scientist | 0.00          |
| 20            | 120         | Electrician    | 0.00          |
| 21            | 125         | Data Scientist | 0.00          |
| 22            | 130         | Electrician    | 0.00          |
| 23            | 135         | Data Scientist | 0.00          |
| 24            | 140         | Electrician    | 0.00          |
| 25            | 145         | Data Scientist | 0.00          |
| 26            | 150         | Electrician    | 0.00          |
| 27            | 155         | Data Scientist | 0.00          |
| 28            | 160         | Electrician    | 0.00          |
| 29            | 165         | Data Scientist | 0.00          |
| 30            | 170         | Electrician    | 0.00          |
| 31            | 175         | Data Scientist | 0.00          |
| 32            | 180         | Electrician    | 0.00          |
| 33            | 185         | Data Scientist | 0.00          |
| 34            | 190         | Electrician    | 0.00          |
| 35            | 195         | Data Scientist | 0.00          |
| 36            | 200         | Electrician    | 0.00          |
| 37            | 205         | Data Scientist | 0.00          |
| 38            | 210         | Electrician    | 0.00          |
| 39            | 215         | Data Scientist | 0.00          |
| 40            | 220         | Electrician    | 0.00          |
| 41            | 225         | Data Scientist | 0.00          |
| 42            | 230         | Electrician    | 0.00          |
| 43            | 235         | Data Scientist | 0.00          |
| 44            | 240         | Electrician    | 0.00          |
| 45            | 245         | Data Scientist | 0.00          |
| 46            | 250         | Electrician    | 0.00          |
| 47            | 255         | Data Scientist | 0.00          |
| 48            | 260         | Electrician    | 0.00          |
| 49            | 265         | Data Scientist | 0.00          |
| 50            | 270         | Electrician    | 0.00          |
| 51            | 275         | Data Scientist | 0.00          |
| 52            | 280         | Electrician    | 0.00          |
| 53            | 285         | Data Scientist | 0.00          |
| 54            | 290         | Electrician    | 0.00          |
| 55            | 295         | Data Scientist | 0.00          |
| 56            | 300         | Electrician    | 0.00          |
| 57            | 305         | Data Scientist | 0.00          |
| 58            | 310         | Electrician    | 0.00          |
| 59            | 315         | Data Scientist | 0.00          |
| 60            | 320         | Electrician    | 0.00          |
| 61            | 325         | Data Scientist | 0.00          |
| 62            | 330         | Electrician    | 0.00          |
| 63            | 335         | Data Scientist | 0.00          |
| 64            | 340         | Electrician    | 0.00          |
| 65            | 345         | Data Scientist | 0.00          |
| 66            | 350         | Electrician    | 0.00          |
| 67            | 355         | Data Scientist | 0.00          |
| 68            | 360         | Electrician    | 0.00          |
| 69            | 365         | Data Scientist | 0.00          |
| 70            | 370         | Electrician    | 0.00          |
| 71            | 375         | Data Scientist | 0.00          |
| 72            | 380         | Electrician    | 0.00          |
| 73            | 385         | Data Scientist | 0.00          |
| 74            | 390         | Electrician    | 0.00          |
| 75            | 395         | Data Scientist | 0.00          |
| 76            | 400         | Electrician    | 0.00          |
| 77            | 405         | Data Scientist | 0.00          |
| 78            | 410         | Electrician    | 0.00          |
| 79            | 415         | Data Scientist | 0.00          |
| 80            | 420         | Electrician    | 0.00          |
| 81            | 425         | Data Scientist | 0.00          |
| 82            | 430         | Electrician    | 0.00          |
| 83            | 435         | Data Scientist | 0.00          |
| 84            | 440         | Electrician    | 0.00          |
| 85            | 445         | Data Scientist | 0.00          |
| 86            | 450         | Electrician    | 0.00          |
| 87            | 455         | Data Scientist | 0.00          |
| 88            | 460         | Electrician    | 0.00          |
| 89            | 465         | Data Scientist | 0.00          |
| 90            | 470         | Electrician    | 0.00          |
| 91            | 475         | Data Scientist | 0.00          |
| 92            | 480         | Electrician    | 0.00          |
| 93            | 485         | Data Scientist | 0.00          |
| 94            | 490         | Electrician    | 0.00          |
| 95            | 495         | Data Scientist | 0.00          |
| 96            | 500         | Electrician    | 0.00          |
| 97            | 505         | Data Scientist | 0.00          |
| 98            | 510         | Electrician    | 0.00          |
| 99            | 515         | Data Scientist | 0.00          |
| 100           | 520         | Electrician    | 0.00          |
| 101           | 525         | Data Scientist | 0.00          |
| 102           | 530         | Electrician    | 0.00          |
| 103           | 535         | Data Scientist | 0.00          |
| 104           | 540         | Electrician    | 0.00          |
| 105           | 545         | Data Scientist | 0.00          |
| 106           | 550         | Electrician    | 0.00          |
| 107           | 555         | Data Scientist | 0.00          |
| 108           | 560         | Electrician    | 0.00          |
| 109           | 565         | Data Scientist | 0.00          |
| 110           | 570         | Electrician    | 0.00          |
| 111           | 575         | Data Scientist | 0.00          |
| 112           | 580         | Electrician    | 0.00          |
| 113           | 585         | Data Scientist | 0.00          |
| 114           | 590         | Electrician    | 0.00          |
| 115           | 595         | Data Scientist | 0.00          |
| 116           | 600         | Electrician    | 0.00          |
| 117           | 605         | Data Scientist | 0.00          |
| 118           | 610         | Electrician    | 0.00          |
| 119           | 615         | Data Scientist | 0.00          |
| 120           | 620         | Electrician    | 0.00          |
| 121           | 625         | Data Scientist | 0.00          |
| 122           | 630         | Electrician    | 0.00          |
| 123           | 635         | Data Scientist | 0.00          |
| 124           | 640         | Electrician    | 0.00          |
| 125           | 645         | Data Scientist | 0.00          |
| 126           | 650         | Electrician    | 0.00          |
| 127           | 655         | Data Scientist | 0.00          |
| 128           | 660         | Electrician    | 0.00          |
| 129           | 665         | Data Scientist | 0.00          |
| 130           | 670         | Electrician    | 0.00          |
| 131           | 675         | Data Scientist | 0.00          |
| 132           | 680         | Electrician    | 0.00          |
| 133           | 685         | Data Scientist | 0.00          |
| 134           | 690         | Electrician    | 0.00          |
| 135           | 695         | Data Scientist | 0.00          |
| 136           | 700         | Electrician    | 0.00          |
| 137           | 705         | Data Scientist | 0.00          |
| 138           | 710         | Electrician    | 0.00          |
| 139           | 715         | Data Scientist | 0.00          |
| 140           | 720         | Electrician    | 0.00          |
| 141           | 725         | Data Scientist | 0.00          |
| 142           | 730         | Electrician    | 0.00          |
| 143           | 735         | Data Scientist | 0.00          |
| 144           | 740         | Electrician    | 0.00          |
| 145           | 745         | Data Scientist | 0.00          |
| 146           | 750         | Electrician    | 0.00          |
| 147           | 755         | Data Scientist | 0.00          |
| 148           | 760         | Electrician    | 0.00          |
| 149           | 765         | Data Scientist | 0.00          |
| 150           | 770         | Electrician    | 0.00          |
| 151           | 775         | Data Scientist | 0.00          |
| 152           | 780         | Electrician    | 0.00          |
| 153           | 785         | Data Scientist | 0.00          |
| 154           | 790         | Electrician    | 0.00          |
| 155           | 795         | Data Scientist | 0.00          |
| 156           | 800         | Electrician    | 0.00          |
| 157           | 805         | Data Scientist | 0.00          |
| 158           | 810         | Electrician    | 0.00          |
| 159           | 815         | Data Scientist | 0.00          |
| 160           | 820         | Electrician    | 0.00          |
| 161           | 825         | Data Scientist | 0.00          |
| 162           | 830         | Electrician    | 0.00          |
| 163           | 835         | Data Scientist | 0.00          |
| 164           | 840         | Electrician    | 0.00          |
| 165           | 845         | Data Scientist | 0.00          |
| 166           | 850         | Electrician    | 0.00          |
| 167           | 855         | Data Scientist | 0.00          |
| 168           | 860         | Electrician    | 0.00          |
| 169           | 865         | Data Scientist | 0.00          |
| 170           | 870         | Electrician    | 0.00          |
| 171           | 875         | Data Scientist | 0.00          |
| 172           | 880         | Electrician    | 0.00          |
| 173           | 885         | Data Scientist | 0.00          |
| 174           | 890         | Electrician    | 0.00          |
| 175           | 895         | Data Scientist | 0.00          |
| 176           | 900         | Electrician    | 0.00          |
| 177           | 905         | Data Scientist | 0.00          |
| 178           | 910         | Electrician    | 0.00          |
| 179           | 915         | Data Scientist | 0.00          |
| 180           | 920         | Electrician    | 0.00          |
| 181           | 925         | Data Scientist | 0.00          |
| 182           | 930         | Electrician    | 0.00          |
| 183           | 935         | Data Scientist | 0.00          |
| 184           | 940         | Electrician    | 0.00          |
| 185           | 945         | Data Scientist | 0.00          |
| 186           | 950         | Electrician    | 0.00          |
| 187           | 955         | Data Scientist | 0.00          |
| 188           | 960         | Electrician    | 0.00          |
| 189           | 965         | Data Scientist | 0.00          |
| 190           | 970         | Electrician    | 0.00          |
| 191           | 975         | Data Scientist | 0.00          |
| 192           | 980         | Electrician    | 0.00          |
| 193           | 985         | Data Scientist | 0.00          |
| 194           | 990         | Electrician    | 0.00          |
| 195           | 995         | Data Scientist | 0.00          |
| 196           | 1000        | Electrician    | 0.00          |
| 197           | 1005        | Data Scientist | 0.00          |
| 198           | 1010        | Electrician    | 0.00          |
| 199           | 1015        | Data Scientist | 0.00          |
| 200           | 1020        | Electrician    | 0.00          |
| 201           | 1025        | Data Scientist | 0.00          |
| 202           | 1030        | Electrician    | 0.00          |
| 203           | 1035        | Data Scientist | 0.00          |
| 204           | 1040        | Electrician    | 0.00          |
| 205           | 1045        | Data Scientist | 0.00          |
| 206           | 1050        | Electrician    | 0.00          |
| 207           | 1055        | Data Scientist | 0.00          |
| 208           | 1060        | Electrician    | 0.00          |
| 209           | 1065        | Data Scientist | 0.00          |
| 210           | 1070        | Electrician    | 0.00          |
| 211           | 1075        | Data Scientist | 0.00          |
| 212           | 1080        | Electrician    | 0.00          |
| 213           | 1085        | Data Scientist | 0.00          |
| 214           | 1090        | Electrician    | 0.00          |
| 215           | 1095        | Data Scientist | 0.00          |
| 216           | 1100        | Electrician    | 0.00          |
| 217           | 1105        | Data Scientist | 0.00          |
| 218           | 1110        | Electrician    | 0.00          |
| 219           | 1115        | Data Scientist | 0.00          |
| 220           | 1120        | Electrician    | 0.00          |
| 221           | 1125        | Data Scientist | 0.00          |
| 222           | 1130        | Electrician    | 0.00          |
| 223           | 1135        | Data Scientist | 0.00          |
| 224           | 1140        | Electrician    | 0.00          |
| 225           | 1145        | Data Scientist | 0.00          |
| 226           | 1150        | Electrician    | 0.00          |
| 227           | 1155        | Data Scientist | 0.00          |
| 228           | 1160        | Electrician    | 0.00          |
| 229           | 1165        | Data Scientist | 0.00          |
| 230           | 1170        | Electrician    | 0.00          |
| 231           | 1175        | Data Scientist | 0.00          |
| 232           | 1180        | Electrician    | 0.00          |
| 233           | 1185        | Data Scientist | 0.00          |
| 234           | 1190        | Electrician    | 0.00          |
| 235           | 1195        | Data Scientist | 0.00          |
| 236           | 1200        | Electrician    | 0.00          |
| 237           | 1205        | Data Scientist | 0.00          |
| 238           | 1210        | Electrician    | 0.00          |
| 239           | 1215        | Data Scientist | 0.00          |
| 240           | 1220        | Electrician    | 0.00          |
| 241           | 1225        | Data Scientist | 0.00          |
| 242           | 1230        | Electrician    | 0.00          |
| 243           | 1235        | Data Scientist | 0.00          |
| 244           | 1240        | Electrician    | 0.00          |
| 245           | 1245        | Data Scientist | 0.00          |
| 246           | 1250        | Electrician    | 0.00          |
| 247           | 1255        | Data Scientist | 0.00          |
| 248           | 1260        | Electrician    | 0.00          |
| 249           | 1265        | Data Scientist | 0.00          |
| 250           | 1270        | Electrician    | 0.00          |
| 251           | 1275        | Data Scientist | 0.00          |
| 252           | 1280        | Electrician    | 0.00          |
| 253           | 1285        | Data Scientist | 0.00          |
| 254           | 1290        | Electrician    | 0.00          |
| 255           | 1295        | Data Scientist | 0.00          |
| 256           | 1300        | Electrician    | 0.00          |
| 257           | 1305        | Data Scientist | 0.00          |
| 258           | 1310        | Electrician    | 0.00          |
| 259           | 1315        | Data Scientist | 0.00          |
| 260           | 1320        | Electrician    | 0.00          |
| 261           | 1325        | Data Scientist | 0.00          |
| 262           | 1330        | Electrician    | 0.00          |
| 263           | 1335        | Data Scientist | 0.00          |
| 264           | 1340        | Electrician    | 0.00          |
| 265           | 1345        | Data Scientist | 0.00          |
| 266           | 1350        | Electrician    | 0.00          |
| 267           | 1355        | Data Scientist | 0.00          |
| 268           | 1360        | Electrician    | 0.00          |
| 269           | 1365        | Data Scientist | 0.00          |
| 270           | 1370        | Electrician    | 0.00          |
| 271           | 1375        | Data Scientist | 0.00          |
| 272           | 1380        | Electrician    | 0.00          |
| 273           | 1385        | Data Scientist | 0.00          |
| 274           | 1390        | Electrician    | 0.00          |
| 275           | 1395        | Data Scientist | 0.00          |
| 276           | 1400        | Electrician    | 0.00          |
| 277           | 1405        | Data Scientist | 0.00          |
| 278           | 1410        | Electrician    | 0.00          |
| 279           | 1415        | Data Scientist | 0.00          |
| 280           | 1420        | Electrician    | 0.00          |
| 281           | 1425        | Data Scientist | 0.00          |
| 282           | 1430        | Electrician    | 0.00          |
| 283           | 1435        | Data Scientist | 0.00          |
| 284           | 1440        | Electrician    | 0.00          |
| 285           | 1445        | Data Scientist | 0.00          |
| 286           | 1450        | Electrician    | 0.00          |
| 287           | 1455        | Data Scientist | 0.00          |
| 288           | 1460        | Electrician    | 0.00          |
| 289           | 1465        | Data Scientist | 0.00          |
| 290           | 1470        | Electrician    | 0.00          |
| 291           | 1475        | Data Scientist | 0.00          |
| 292           | 1480        | Electrician    | 0.00          |
| 293           | 1485        | Data Scientist | 0.00          |
| 294           | 1490        | Electrician    | 0.00          |
| 295           | 1495        | Data Scientist | 0.00          |
| 296           | 1500        | Electrician    | 0.00          |
| 297           | 1505        | Data Scientist | 0.00          |
| 298           | 1510        | Electrician    | 0.00          |
| 299           | 1515        | Data Scientist | 0.00          |
| 300           | 1520        | Electrician    | 0.00          |
| 301           | 1525        | Data Scientist | 0.00          |
| 302           | 1530        | Electrician    | 0.00          |
| 303           | 1535        | Data Scientist | 0.00          |
| 304           | 1540        | Electrician    | 0.00          |
| 305           | 1545        | Data Scientist | 0.00          |
| 306           | 1550        | Electrician    | 0.00          |
| 307           | 1555        | Data Scientist | 0.00          |
| 308           | 1560        | Electrician    | 0.00          |
| 309           | 1565        | Data Scientist | 0.00          |
| 310           | 1570        | Electrician    | 0.00          |
| 311           | 1575        | Data Scientist | 0.00          |
| 312           | 1580        | Electrician    | 0.00          |
| 313           | 1585        | Data Scientist | 0.00          |
| 314           | 1590        | Electrician    | 0.00          |
| 315           | 1595        | Data Scientist | 0.00          |
| 316           | 1600        | Electrician    | 0.00          |
| 317           | 1605        | Data Scientist | 0.00          |
| 318           | 1610        | Electrician    | 0.00          |
| 319           | 1615        | Data Scientist | 0.00          |
| 320           | 1620        | Electrician    | 0.00          |
| 321           | 1625        | Data Scientist | 0.00          |
| 322           | 1630        | Electrician    | 0.00          |
| 323           | 1635        | Data Scientist | 0.00          |
| 324           | 1640        | Electrician    | 0.00          |
| 325           | 1645        | Data Scientist | 0.00          |
| 326           | 1650        | Electrician    | 0.00          |
| 327           | 1655        | Data Scientist | 0.00          |
| 328           | 1660        | Electrician    | 0.00          |
| 329           | 1665        | Data Scientist | 0.00          |
| 330           | 1670        | Electrician    | 0.00          |
| 331           | 1675        | Data Scientist | 0.00          |
| 332           | 1680        | Electrician    | 0.00          |
| 333           | 1685        | Data Scientist | 0.00          |
| 334           | 1690        | Electrician    | 0.00          |
| 335           | 1695        | Data Scientist | 0.00          |
| 336           | 1700        | Electrician    | 0.00          |
| 337           | 1705        | Data Scientist | 0.00          |
| 338           | 1710        | Electrician    | 0.00          |
| 339           | 1715        | Data Scientist | 0.00          |
| 340           | 1720        | Electrician    | 0.            |

To build a Naïve Bayesian Classifier, you must collect the following statistics:

- Probability of all class labels—Probability of purchasing a product and probability of not purchasing a product. From the all data available in the training set, you determine  $P(\text{Yes}) = 0.39$  and  $P(\text{No}) = 0.61$ .
  - In the training set, there are several attributes: Income Tier, Occupation, and Age Tier. For each attribute and its possible values, you must compute the conditional probabilities of a customer purchasing a product. For example, relative to the Occupation attribute, you must compute  $P(\text{Electrician}|\text{Yes})$ ,  $P(\text{Electrician}|\text{No})$ ,  $P(\text{Data Scientist}|\text{Yes})$ ,  $P(\text{Data Scientist}|\text{No})$ , and so on.

## Naïve Bayesian classifier example, cont.

### Naïve Bayesian classifier example, cont.

- Given applicant attributes of A = (Age 30–40, Occupation Electrician, Income 80–120)
- Since  $P(\text{No}|A) > P(\text{Yes}|A)$ , assign the label No, the customer will not purchase.  
 $P(\text{Yes}|A) \sim (0.21 * 0.42 * 0.28) / 0.39 = 0.009$   
 $P(\text{No}|A) \sim (0.36 * 0.22 * 0.40) / 0.81 = 0.019$

| a <sub>i</sub> | C <sub>i</sub> | P(a <sub>i</sub> ; C <sub>i</sub> ) |
|----------------|----------------|-------------------------------------|
| 30-40          | Yes            | 0.21                                |
| 30-40          | No             | 0.36                                |
| Electrician    | Yes            | 0.42                                |
| Electrician    | No             | 0.22                                |
| 80-120         | Yes            | 0.28                                |
| 80-120         | No             | 0.40                                |

Data Science and Big Data Analytics v2

Dell EMC

Here, you have an example of a customer who belongs to age group 30 to 40, is an Electrician, and falls in the income bracket of 80 to 120. How should you classify this customer, as a customer who will buy the product or not buy?

Having built the classifier with the training set, you find  $P(\text{Yes}|A)$  is equal to 0.009—see the computation shown here—and  $P(\text{No}|A)$  is 0.019. Since  $P(\text{No}|A)$  is the maximum of the two probability scores, assign the customer to the group not buying the product.

The score is only proportional to the probability. It does not equal the probability, because you have not included the denominator. However, both formulas have the same denominator, so it is not necessary to calculate it to know which quantity is bigger.

Notice, though, how small in magnitude these scores are. When you are looking at problems with many attributes, or attributes with a very high number of levels, these values can become very small in magnitude.

## Naïve Bayesian implementation considerations

**Naïve Bayesian implementation considerations**

- Numerical underflow
  - Resulting from multiplying several probabilities near zero
  - Preventable by computing the logarithm of the products
- Zero probabilities due to unobserved attribute/classifier pairs
  - Resulting from rare events
  - Handled by smoothing—adjusting each probability by a small amount
- Assign the classifier label,  $C_i$ , that maximizes the value of

$$\left( \sum_{j=1}^w \log P'(a_j | C_i) \right) + \log P(C_i)$$

where  $i = 1, 2, \dots, n$  and  
 $P'$  denotes the adjusted probabilities

Multiplying several probability values, each possibly close to zero, invariably leads to the problem of numerical underflow. So, an important implementation guideline is to compute the logarithm of the product of the probabilities, which is equivalent to the summation of the logarithm of the probabilities. Although the risk of underflow may increase as the number of attributes increase, the use of logarithms should be applied regardless of the number of attribute dimensions.

Further, to address the possibility of probabilities equal to zero, smoothing techniques can be employed to adjust the probabilities to ensure nonzero values. Applying a smoothing technique assigns a small nonzero probability to rare events not included in the training dataset. Also, the smoothing addresses the possibility of taking the logarithm of zero.

The R implementation of Naïve Bayes incorporates the smoothing directly into the probability tables. Essentially, the Laplace smoothing that R uses adds one, or a small value, to every count. For example, if you have 100 "good" customers, and 20 of them rent their housing, the "raw"  $P(\text{rent} | \text{good})$  equals  $20/100 = 0.2$ . With Laplace smoothing adding one to the counts, the calculation would be  $P(\text{rent} | \text{good}) \sim (20 + 1)/(100+3) = 0.20388$ , where there are three possible values for housing—own, rent, for free.

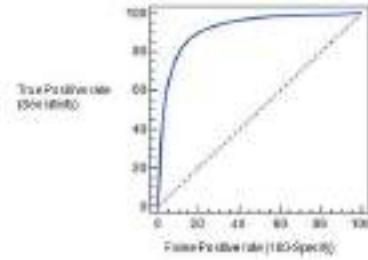
Fortunately, the uses of the logarithms and the smoothing techniques are already implemented in standard software packages for Naïve Bayes Classifiers. However,

if for performance reasons, the Naïve Bayes Classifier algorithm must be coded directly into an application, these considerations should be implemented.

## Diagnostics

### Diagnostics

- Hold-out data.
  - How well does the model classify new instances?
- Cross-validation
- ROC curve/AUC
- Confusion Matrix



14

DATA SCIENCE

DELL INC.

The diagnostics you used in regression can be used to validate the effectiveness of the model you built. The technique of using the hold-out data and performing N-fold cross validations and using the ROC/Area Under the Curve methods can be deployed with Naïve Bayesian Classifier, as well.

## Naïve Bayesian classifier—reasons to choose (+) and cautions (-)

| Naïve Bayesian classifier—reasons to choose (+) and cautions (-)   |   |
|--|---|
| Reasons to choose (+)  | Cautions (-)  |
| Handles missing values quite well  | Numerical variables must be discrete, categorical, binary                       |
| Robust to irrelevant variables   | Sensitive to correlated variables<br>Double-counting                            |
| Easy to implement  | Not good for estimating probabilities<br>Sensitivity to class label frequencies |
| Easy to score data   |   |
| Resistant to overfitting   |   |
| Computationally efficient:<br>Handles very high-dimensional problems<br>Handles categorical variables with many levels |   |

The reasons to choose (+) and cautions (-) of the Naïve Bayesian Classifier are listed. Unlike Logistic regression, Naïve Bayesian Classifier handles missing values well. It is also robust to irrelevant variables—irrelevant variables are distributed among all the classes, and their effects are not pronounced.

The model is easy to implement, and a basic version can be implemented in the lab without using any packages. Scoring data, predicting, is simple, and the model is resistant to overfitting. Overfitting refers to fitting training data so well that you fit the idiosyncrasies, such as the data that is not relevant in characterizing the data. It is computationally efficient and handles high-dimensional problems efficiently. Unlike logistic regression, Naïve Bayesian Classifier handles categorical variables with many levels.

The cautions (-) are that it is sensitive to correlated variables, as the algorithm double counts the effect of the correlated variables. For example, people with low income tend to default and people with low credit tend to default. It is also true that people with low income tend to have low credit. If you try to score "default" with both low income and low credit as variables, you see the double-counting effect in your model output and in the scoring.

Though the probabilities are provided as an output of the scored data, Naïve

## Lesson: Naïve Bayes

Bayesian Classifier is not very reliable for the probability estimation and should be used for class label assignments only. Naïve Bayesian Classifier in its simple form is used only with categorical variables, and any continuous variables should be rendered discrete into intervals. The lab provides more information on this idea. However, it is not necessary to have the continuous variables as discrete, and several standard implementations can handle continuous variables, as well.

## Check your knowledge

**Check your knowledge**

1. Consider the following training dataset:
  - Apply the Naïve Bayesian Classifier to this dataset and compute the probability score for  $P(y = 1|X)$  for  $X = (1,0,0)$
  - Show your work
2. List some prominent use cases of the Naïve Bayesian Classifier.
3. What gives the Naïve Bayesian Classifier the advantage of being computationally inexpensive?
4. Why should you use logs in the probability scoring calculations?

| X1 | X2 | X3 | Y |
|----|----|----|---|
| 1  | 1  | 1  | 0 |
| 1  | 1  | 0  | 0 |
| 0  | 0  | 0  | 0 |
| 0  | 1  | 0  | 1 |
| 1  | 0  | 1  | 1 |
| 0  | 1  | 1  | 1 |

Training Dataset

Dell EMC

## Check your knowledge



### Discussion

## Question / Discussion Topic:

1. Consider the following training dataset:
  - Apply the Naïve Bayesian Classifier to this dataset and compute the probability score for  $P(y = 1|X)$  for  $X = (1,0,0)$
  - Show your work
2. List some prominent use cases of the Naïve Bayesian Classifier.
3. What gives the Naïve Bayesian Classifier the advantage of being computationally inexpensive?
4. Why should you use logs in the probability scoring calculations?

**Discussion Notes:**

## Check your knowledge, cont.

### Check your knowledge, cont.

1. Consider the following dataset with two input features, temperature and season:
- What is the Naïve Bayesian assumption?
  - Is the Naïve Bayesian assumption satisfied for this problem?

| Temperature | Season | Electricity Usage |
|-------------|--------|-------------------|
| -10 to 50 F | Winter | High              |
| 50 to 70 F  | Winter | Low               |
| 70 to 85 F  | Summer | Low               |
| 85 to 110 F | Summer | High              |

## Check your knowledge



### Discussion

### Question / Discussion Topic:

- Consider the following dataset with two input features, temperature and season:
  - What is the Naïve Bayesian assumption?
  - Is the Naïve Bayesian assumption satisfied for this problem?

### Discussion Notes:

## Naïve Bayesian classifiers—summary

### Naïve Bayesian classifiers—summary

During this lesson, the following topics were covered:

- Naïve Bayesian Classifier
- Theoretical foundations of the classifier
- Use cases
- Evaluating the effectiveness of the classifier
- The reasons to choose (+) and cautions (-) with the use of the classifier



Source: DELL INC.

DELL INC.

This lesson covered these topics. Take a moment to review them.

## Lesson: Decision trees

Introduction

Lesson: Decision trees



## Decision Trees

### Decision Trees

During this lesson, the following topics are covered:

- Overview of Decision Tree classifier
- General algorithm for Decision Trees
- Decision Tree use cases
- Entropy, Information gain
- Reasons to choose (+) and cautions (-) of Decision Tree classifier
- Classifier methods and conditions in which they are best suited

The topics covered in this lesson are listed.

## Decision Tree classifier—what is it?

### Decision Tree classifier—what is it?

- Used for classification:
  - Returns probability scores of class membership
    - Well-calibrated, as is logistic regression
    - Assigns label based on highest scoring class
    - Some Decision Tree algorithms return simply the most likely class
  - Regression Trees: a variation for regression
    - Returns average value at every node
    - Predictions can be discontinuous at the decision boundaries
- Input variables can be continuous or discrete
- Output:
  - This output is a tree that describes the decision flow.
  - Leaf nodes return either a probability score or simply a classification.
  - Trees can be converted to a set of "decision rules."
    - "If income < \$50,000 AND mortgage\_amt > \$100K THEN default=T with 75% probability."

140 / 800 UNSTRUCTURED DATA

DELL EMC

Decision Trees are a flexible method commonly deployed in data mining applications. This lesson presents Decision Trees used for classification problems.

There are two types of trees: Classification Trees and Regression—or, Prediction—Trees.

- Classification Trees are used to segment observations into more homogeneous groups—assign class labels. They usually apply to outcomes that are binary or categorical in nature.
- Regression Trees are variations of regression. And, what is returned in each node is the average value at each node—type of a step function with which the average value can be computed. Regression trees can be applied to outcomes that are continuous, such as account spend or personal income.

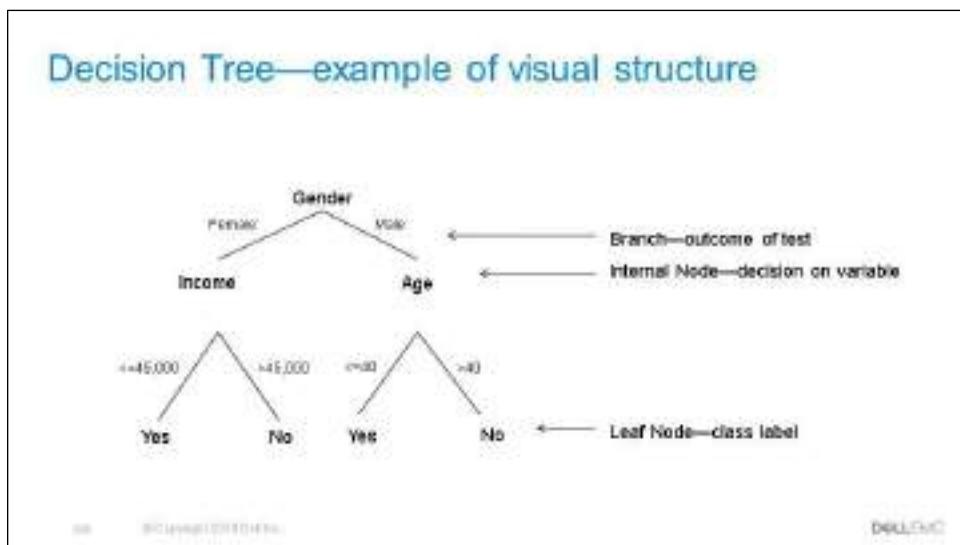
The input values can be continuous or discrete. Decision Tree models output a tree that describes the decision flow. The leaf nodes return class labels and, in some implementations, they also return the probability scores. In theory, the tree can be converted into decision rules such as the example shown here.

Decision Trees are a popular method because they can be applied to various situations. The rules of classification are straightforward, and the results can easily be presented visually. Further, because the result is a series of logical "if-then"

## Lesson: Decision trees

statements, there is no underlying assumption of a linear—or nonlinear—relationship between the predictor variables and the dependent variable.

## Decision Tree—example of visual structure



Decision Trees are typically depicted in a flow-chart-like manner.

**Branches** refer to the outcome of a decision; the connecting lines here represent branches.

When the decision is numerical, the "greater than" branch is usually shown on the right and "less than" on the left.

Depending on the nature of the variable, it may be necessary to include an "equal to" component on one branch.

**Internal Nodes** are the decision or test points. Each refers to a single variable or attribute.

In the example here, the outcomes are binary, although there could be more than two branches stemming from an internal node. For example, if the variable was categorical and had three choices, you might need a branch for each choice.

The **Leaf Nodes** are at the end of the last branch on the tree. These nodes represent the outcome of all the prior decisions. The leaf nodes are the class labels, or the segment in which all observations that follow the path to the leaf would be placed.

## Decision Tree classifier—use cases

### Decision Tree classifier—use cases

- When a series of questions (yes/no) are answered to arrive at a classification
  - Biological species classification
  - Checklist of symptoms during a doctor's evaluation of a patient
- When "if-then" conditions are preferred to linear models
  - Customer segmentation to predict response rates
  - Financial decisions such as loan approval
  - Fraud detection
- Short Decision Trees are the most popular "weak learner" in ensemble learning techniques

000

DATA SCIENCE

DELL INC.

An example of Decision Trees in practice is the method for classifying biological species. A series of questions (yes/no) are answered to arrive at a classification.

Another example is a checklist of symptoms during a doctor's evaluation of a patient. People mentally perform these types of analysis frequently when assessing a situation.

Other use cases can be customer segmentation to better predict response rates to marketing and promotions. Computers can be "taught" to evaluate a series of criteria and automatically approve or deny an application for a loan. For loan approval, computers can use the logical "if-then" statements to predict whether the customer will default on the loan. For customers with a clear—strong—outcome, no human interaction is necessary. For observations that may not generate a clear response, a human is needed for the decision.

In short Decision Trees, you have a limited the number of splits. These splits are often used as components, called "weak learners" or "base learners," in ensemble techniques—a set of predictive models that will all vote, and you take decisions based on the combination of the votes—such as Random forests, or bagging and boosting. These techniques are beyond the scope for this class. The simplest of the short trees are decision stumps: Decision Trees with one internal node, the root, which is immediately connected to the terminal nodes. A decision stump makes a prediction based on the value of just a single input feature.

## Example—credit risk problem



For the people with good credit, starting at the top of the tree, the probability is 70 percent; 700 out of 1000 people have good credit. The process dictates that the income amounts be split into two groups.

One group contains income amounts of less than 35 K, and the other includes income amounts greater than or equal to 35 K.

Compute the probability of good credit at the second node, and you find that the values in the income group with income of more than 35 K have a lower probability, 57 percent, of being in the high-risk category. You split this group on credit history, this time. One group includes people with a good credit history, and the other includes people with a bad or unknown credit history. The group with a good credit history has a low probability of high risk—37 percent. The group with an unknown or bad credit history has a high percentage of being high-risk customers—85 percent.

The probability of good credit at the other node shows that people in the income group with income of less than 35 K have a higher probability—83 percent—of being in the high-risk category. Now, you split this group on credit history. One group's members have a good credit history, and the other's have a bad or unknown credit history. The group with a good credit history has a low probability of high risk—41 percent. The group with an unknown or bad credit history is then split again, based on the debt of the customer. Members of the group with a high amount of debt have a high probability of being high-risk customers—95 percent.

## Lesson: Decision trees

And, members of the group with a low amount of debt have a lower probability of being high risk—43 percent.

Decision Trees are greedy algorithms. They take decisions based on what is available at that moment. After a bad decision is taken, it is propagated all the way down. An ensemble technique may randomize the splitting, or even randomize data, and come up with multiple tree structures. Then, by looking at the average of the nodes in all the trees, it assigns class labels or probability values.

## General algorithm

### General algorithm

- To construct tree T from training set S
  - If all examples in S belong to some class in C, or S is sufficiently "pure", then make a leaf labeled C.
  - Otherwise:
    - Select the "most informative" attribute A.
    - Partition S according to A's values.
    - Recursively construct subtrees T1, T2, and so on, for the subsets of S.
- The details vary according to the specific algorithm—CART, ID3, C4.5—but the general idea is the same.

30

DATA SCIENCE

DELL INC.

Now, describe the general algorithm. The objective is to construct a tree T from a training set S. If all the examples in S belong to some class "C"—good\_credit, for example—or if S is sufficiently "pure," you make a leaf labeled "C". In this case, node p(credit\_good) is 70 percent pure.

Otherwise, you select another attribute considered as the most informative, such as savings, housing, and so on. Then, partition S according to A's values. Something similar to the process explained in the previous slide. Construct subtrees T1, T2, and so on, or the subsets of S recursively, until:

- You have all of the nodes as pure as is needed or
- You cannot split further as per your specifications or
- You reach any other stopping criteria specified.

Several algorithms implement Decision Trees, and the methods of tree construction vary with each one of them. CART, ID3 and C4.5 are some of the popular algorithms.

## Step 1—Pick most informative attribute

### Step 1—Pick most informative attribute

Entropy-based methods are one common way:

$$H = - \sum_c p(c) \log_2 p(c)$$

$H=0$  if  $p(c) = 0$  or  $1$  for any class.

So, for binary classification,  $H=0$  is a pure node.

$H$  is maximum when all classes are equally probable.

For binary classification,  $H=1$  when classes are 50/50.

The first step is to pick the most informative attribute. There are many ways to do it. For this example, detail Entropy-based methods.

Let  $p(c)$  be the probability of a given class.  $H$ , as defined by the formula shown here, has a value 0 if  $p(c)$  is 0 or 1. So, for binary classification,  $H=0$  means it is a pure node.  $H$  is maximum when all classes are equally probable. If the probability of classes is 50/50, then  $H=1$ , maximum entropy.

## Step 1—Pick most informative attribute—conditional entropy

### Step 1—Pick most informative attribute—conditional entropy

$$H_{attr} = - \sum_v p(v) \sum_c p(c|v) \log_2 p(c|v)$$

The weighted sum of the class entropies for each value of the attribute.

In English, attribute values—homeowner vs. renter—give more information about class membership.

"Homeowners are more likely to have good credit than renters."

Conditional entropy should be lower than unconditioned entropy.

Continuing with step 1 you now find the conditional entropy, which is the weighted sum of class entropies for each value of the attribute.

Suppose that you choose the attribute "Housing". You have three levels for this attribute—free, rent, and own. Intuitively, you can say that homeowners are more likely to have better credit than renters. So, the attribute value "Housing" gives more information about the class membership for credit\_good. The conditional entropy of the attribute "Housing" should be lower than the base entropy.

At worst, in the case where the attribute is uncorrelated with the class label, the conditional entropy is the same as the unconditioned entropy.

## Step 1—which attribute is best classifier?

### Step 1—which attribute is best classifier?

A statistical property called information gain, measures how well a given attribute separates the training examples.

Information gain uses the notion of entropy, commonly used in information theory.

$$\text{InfoGain}_{\text{attr}} = H - H_{\text{attr}}$$

Information gain = expected reduction of entropy

$H$  is entropy at the first node, and  $H(\text{leaf})$  is entropy of the leaf nodes.

Here is an example of two separate attributes from a class with 29 positive and 35 negative responses.



CC BY-SA 3.0 Dell Inc.

Dell EMC

Information Gain is defined as the difference between the base entropy and the conditional entropy of the attribute.

So, the most informative attribute is the attribute with most information gain.

Remember, this scenario is just an example. There are other information/purity measures, but InfoGain is a fairly popular one for inducing Decision Trees.

As seen in this example, an attribute with 29 positive and 35 negative responses is split into two parts using the A1 and A2 attribute. Both attributes result in two different leaf nodes. How can you know the information value of each attribute?

## Step 1—which attribute is best classifier? (cont.)

**Step 1—which attribute is best classifier? (cont.)**

**Entropy Calculation at root node**

$$\text{entropy } ((29+35) = -\left(\frac{29}{64}\right)\log_2\left(\frac{29}{64}\right) - \left(\frac{35}{64}\right)\log_2\left(\frac{35}{64}\right) \approx 0.994$$

**Information gain of the two variables**

```

graph TD
    Root((29+35)) --> A1((A1=x1))
    Root --> A2((A2=x1))
    A1 --> L1((E=0.706))
    A1 --> R1((E=0.742))
    A2 --> L2((E=0.937))
    A2 --> R2((E=0.819))
  
```

Information gain calculation for A1:

$$\begin{aligned} I(A1) &= H(0.994) - (0.706 \cdot 0.736 + 0.742 \cdot 0.714) \\ &= 0.296 \\ &= 0.296 / 0.994 \times 100\% = 30\% \end{aligned}$$

Information gain calculation for A2:

$$\begin{aligned} I(A2) &= H(0.994) - (0.937 \cdot 0.518 + 0.819 \cdot 0.264) \\ &= 0.121 \\ &= 0.121 / 0.994 \times 100\% = 12.2\% \end{aligned}$$

© 2018 Dell Inc. All rights reserved.

Dell EMC

Here, you see the entropy calculation for root node.

To calculate the information gain of an attribute, you must first calculate entropy at each node using the same formula that was used for the root node.

After you have the entropy values at each node, you can calculate the information gain using the formula from the previous slide.

## Conditional entropy example

| Conditional entropy example |          |       |       |
|-----------------------------|----------|-------|-------|
|                             | positive | 0%    | 100%  |
| positive                    | 0.108    | 0.718 | 0.100 |
| (0%) housing                | 0.407    | 0.261 | 0.391 |
| (negative) housing          | 0.592    | 0.738 | 0.601 |

$$H(\text{housing}|\text{credit}) = -(0.108 * (0.407 \log_2(0.407) + 0.592 \log_2(0.592)) \\ + 0.718 * (0.261 \log_2(0.261) + 0.738 \log_2(0.738)) \\ + 0.100 * (0.391 \log_2(0.391) + 0.601 \log_2(0.601)) \\ = 0.618$$

Source: https://www.cs.cmu.edu/~mlbook/

Dell EMC

Here is a look at how to compute the conditional entropy of credit class conditioned on housing status.

In the top row of the table are the probabilities of each value. In the next two rows are the probabilities of the class labels conditioned on the housing value.

Each term inside parentheses is the entropy of the class labels within a single housing value.

The conditional entropy is still fairly high; but it is a little less than the unconditioned entropy.

## Steps 2 and 3—partition on selected variable

**Steps 2 and 3—partition on selected variable**

- Step 2: Find the partition with the highest InfoGain.
  - In this example, the selected partition has InfoGain = 0.028.
- Step 3: At each resulting node, repeat Steps 1 and 2.
  - Until node is "pure enough"
- Pure nodes → no information gain by splitting on other attributes

```

graph TD
    Root[Merge] --> L1[Savings <= 100, (100, 500)]
    Root --> R1[Savings > 1000, (no brown savings)]
    L1 --> Leaf1[Merge]
    R1 --> Leaf2[Merge]
  
```

The selected partitioning has InfoGain almost as high as using each savings value as a separate node. And InfoGain happens to be biased to many partitions, so this partition is as informative.

InfoGain can be used with continuous variables, as well; in that case, finding the partition and computing the information gain are the same step.

"Pure enough" usually means that no more information can be gained by splitting on other attributes.

## Diagnostics

### Diagnostics

- Hold-out data
- ROC/AUC
- Confusion Matrix
- FPR/FNR, Precision/Recall
- Do the splits—or the rules—make sense?
  - What does the domain expert say?
- How deep is the tree?
  - Too many layers are prone to overfit.
- Do you get nodes with very few members?
  - Overfit



© 2018 Dell Inc. All rights reserved.

The diagnostics are the same as the one you detailed for Naïve Bayesian classifier. You use the hold-out data /AUC and confusion matrix. There are sanity checks that can be performed, such as validating the decision rules with domain experts and determining if they make sense.

Having too many layers and obtaining nodes with very few members are signs of overfitting.

## Decision Tree classifier— reasons to choose (+) and cautions (-)

| Decision Tree classifier— reasons to choose (+) and cautions (-) |  |
|--|--|
| Reasons to choose (+)  | Cautions (-)   |
| Takes any datatype—numerical, categorical.                       | Tree structure is very sensitive to small changes in the training data.  |
| Handles both numerical and categorical variables.                |  |
| Handles missing values naturally.                                | And it is very easy to overfit.  |
| Handles any number of features in one prediction.                |  |
| Computationally efficient to build.                              | Classical tree methods including CART, decision trees, random forests, etc., can take a long time to build a tree with many nodes. |
| Easy to understand.  | In practice, decision trees can be fairly complex.   |
| Many algorithms can measure importance of variables.             |  |
| Decision rules are easy to understand.                           |  |

Decision Trees take both numerical and categorical variables. They can handle many distinct values such as the zip code in the data. Unlike Naïve Bayes, the Decision Tree method is robust with redundant or correlated variables. Decision Trees handles variables that are nonlinear. Linear/logistic regression computes the value as  $b_1*x_1 + b_2*x_2$ , and so on.

Naïve Bayes also does not do variable interactions, by design. Decision Trees handle variable interactions naturally. Every node in the tree is in some sense an interaction.

Decision Tree algorithms are computationally efficient, and it is easy to score the data. The outputs are easy to understand. Many algorithms return a measure of variable importance. Typically, many software packages provide the information gain from each variable.

In terms of cautions (-), the tree structure is sensitive to small variations in the training data. If you have a large dataset, and you build a Decision Tree on one subset and another Decision Tree on a different subset, the resulting trees can be very different, even though they are from the same dataset. The tree depends on which variable is chosen for the first split. If you get a deep tree, you are probably overfitting as each split reduces the training data for subsequent splits.

## Lesson: Decision trees

If you have redundant variables, Decision Trees ignore them as the algorithm cannot detect any information gain. If their variables are important, and if you split on these variables, you end up with less data with every split.

If you are modeling with logistic regression and you have 500 variables and you really do not know which ones to choose, you can use Decision Trees to determine which variables to select, based on information gain. Then choose those variables for the logistic regression model. Decision Trees can be used to prune redundant variables.

Decision Trees do not naturally handle missing values, though many implementations include a method for dealing with this issue. Even though decision rules are easy to understand, in practice they can be complex.

### References

Hastie, Tibshirani, and Friedman, "The Elements of Statistical Learning"  
Seni and Elder, "Ensemble Methods In Data Mining"

## Which classifier should I try?

| Reason/Question   | Recommended classifier               |
|---|--------------------------------------|
| Do I want a classifier that is easier to understand?                                      | Logistic regression<br>Decision Tree |
| Do I want it to handle how the variables affect the model?                                | Logistic regression<br>Decision Tree |
| Is the dataset high dimensional?  | Random Forest                        |
| Do I expect some of the inputs are correlated?  | Decision Tree<br>Logistic regression |
| Do I expect some of the inputs are irrelevant?  | Decision Tree<br>Random Forest       |
| Are there categorical variables with many levels?   | Random Forest<br>Decision Tree       |
| Are there mixed variable types?   | Decision Tree<br>Logistic regression |
| Is there a lot of noise in the output labels or is the model that will affect the output? | Decision Tree                        |

This list is only advisory. It is a list of things to think about when picking a classifier, based on the reasons to choose (+) and cautions (-).

## Check your knowledge

### Check your knowledge

1. How do you define information gain?
2. For what conditions is the value of entropy at a maximum and when is it at a minimum?
3. List three use cases of decision trees.
4. What are weak learners and how are they used in ensemble methods?
5. Why do you end up with an overfitted model with deep trees and in datasets when you have outcomes that depend on many variables?
6. What classification method would you recommend for the following cases:
  - + High-dimensional data
  - + Data in which outputs are affected by nonlinearity and discontinuity in the inputs



## Check your knowledge



### Discussion

## Question / Discussion Topic:

1. How do you define information gain?
2. For what conditions is the value of entropy at a maximum and when is it at a minimum?
3. List three use cases of decision trees.
4. What are weak learners and how are they used in ensemble methods?
5. Why do you end up with an overfitted model with deep trees and in datasets when you have outcomes that depend on many variables?
6. What classification method would you recommend for the following cases:
  - High-dimensional data

- Data in which outputs are affected by nonlinearity and discontinuity in the inputs

**Discussion Notes:**

## Decision trees—summary

### Decision trees—summary

During this lesson, the following topics were covered:

- Overview of decision tree classifier
- General algorithm for decision trees
- Decision tree use cases
- Entropy, information gain
- Reasons to choose (+) and cautions (-) of decision tree classifier
- Classifier methods and conditions in which they are best suited



This lesson covered these topics. Take a moment to review them.

## Lesson: Time series analysis

Introduction

Lesson: Time series  
analysis

DELL EMC

## Time Series Analysis

### Time Series Analysis

During this lesson, the following topics are covered:

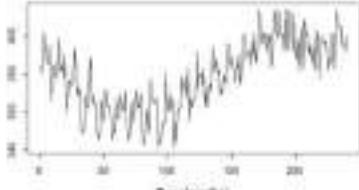
- Time Series Analysis and its applications in forecasting
- Autoregressive Moving Averages (ARMA) and ARIMA Models
- Implementing the Box-Jenkins Methodology using R
- Reasons to choose (+) and cautions (-) with Time Series Analysis

The topics covered in this lesson are listed. ARIMA and Box-Jenkins methodology are explained next.

## Time Series Analysis, cont.

### Time Series Analysis, cont.

- Time Series: Ordered sequence of equally spaced values over time.
- Analysis Goals
  - To identify the internal structure of the time series
  - To forecast future events
    - Example: Based on sales history, what will next December's sales be?



140 150 160 170 180 190 200  
Time (months)

DELL INC.

Businesses perform sales forecasting to look ahead to plan their investments, launch new products, decide when to close or withdraw products, and so on. The sales forecasting process is a critical one for most businesses. Part of the sales forecasting process is to examine the past. How well did the company do in the last few months, or what were the sales in the same time period for the last few years? Time Series Analysis provides a scientific methodology for sales forecasting. **Time Series Analysis** is the analysis of sequential data across equally spaced units of time. Time Series is a basic research methodology in which data for one or more variables is collected for many observations at different time periods.

The main objectives in Time Series Analysis are:

- To understand the underlying structure of the time series by breaking it down to its components.
- Fit a mathematical model and then proceed to forecast the future.

The time periods are equally spaced, and the observations may be either univariate or multivariate. **Univariate** time series are those series where only one variable is measured over time, whereas multivariate time series are those series where multiple variables are measured simultaneously.

## Components of Time Series Analysis

### Components of Time Series Analysis

- A time series may involve the following components:
  - Trend
  - Seasonality
  - Cycles
  - Random
- Method: Box-Jenkins—ARMA

111

DATA SCIENCE

DELL EMC

The internal structure of the data may specify a trend, seasonality, or cycles:

**Trend component**—Trend is a long-term movement in a time series. It is the underlying direction—upward or downward—and rate of change in a time series, when allowance has been made for the other components.

**Seasonal component**—Seasonal fluctuations of known periodicity. It is the component of variation in a time series that depends on the time of the year. It describes any regular fluctuations with a period of less than one year. For example, the costs of various types of fruits and vegetables, and average daily rainfall, all show marked seasonal variation.

**Cyclic component**—Cyclical variations of nonseasonal nature, whose periodicity is unknown.

**Random component**—Random or chaotic values left over when other components of the series—trend, seasonal, and cyclical—have been accounted for.

This lesson primarily presents one analysis methodology, known as the Box-Jenkins method.

## Box-Jenkins method—what is it?

Box-Jenkins method—what is it?

Models historical behavior to forecast the future.

AirPassengers

Time

Applies ARMA

Accounting for the Trends and Seasonality components.

Input: Time Series

Output: Expected future value of the time series.

DATA SOURCE: Box, G.E.P. and Jenkins, G.M. (1970).

DELL INC.

Box-Jenkins methodology was developed by Professors G.E.P. Box and G.M. Jenkins. This methodology enables forecasting with time series data, with both high accuracy and low computational requirements.

The technique may be applied to quickly determine forecasts that are as uncomplicated in form as the simple smoothing methods, or that involve various economic variables. In either case, use of this technique enables efficient utilization of other predictive information contained in the data. It offers assurance of obtaining the highest forecasting accuracy possible in terms of the variables on which the forecast is based.

The input for the model is the trend and seasonality adjusted time series, and the output is the expected future value of the time series.

Box-Jenkins Methodology applies autoregressive moving average ARMA models to find the best fit of a time series to past values of this time series, to make forecasts.

## AR and MA models

### AR and MA models

- Autoregressive process
  - AR(p): Current value of the series depends on its own p previous values.
  - p is the order of the AR process.
- Moving Average process
  - MA(q): Current deviation from mean depends on q previous deviations.
  - q is the order of the MA process.

Box-Jenkins Methodology applies autoregressive moving average ARMA models to find the best fit of a time series to past values of this time series, to make forecasts.

## Time series—use cases

Time series—use cases

Forecast:

- Employment Trend
- Next month's sales
- Tomorrow's stock price
- Hourly power demand



© 2018 Dell Inc. All rights reserved.

The key application of **Time Series Analysis** is in forecasting. Economic and business planning, inventory, and production control of industrial processes are some of the key applications in which Time Series Analysis is deployed.

Time Series data provide useful information about the physical, biological, social, or economic systems generating the time series, such as:

**Economics and Finance:** share prices, profits, imports, exports, stock exchange indexes

**Sociology:** school enrollments, unemployment, crime rate

**Environment:** Amount of pollutants, such as suspended particulate matter (SPM), in the environment

**Meteorology:** Rainfall, temperature, wind speed

**Epidemiology:** Number of flu cases over time

**Medicine:** Blood pressure measurements over time for evaluating drugs to control hypertension

## Modeling time series

### Modeling time series

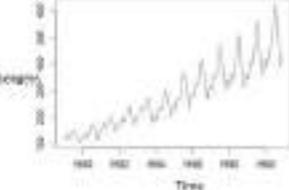
Model the time series as

$$Y_t = T_t + S_t + R_t, \quad t=1, \dots, n.$$

T<sub>t</sub>: Trend term  
Air travel steadily increased over the last few years.

S<sub>t</sub>: The seasonal term  
Air travel fluctuates in a regular pattern over the course of a year.

R<sub>t</sub>: Random component  
To be modeled with ARMA.

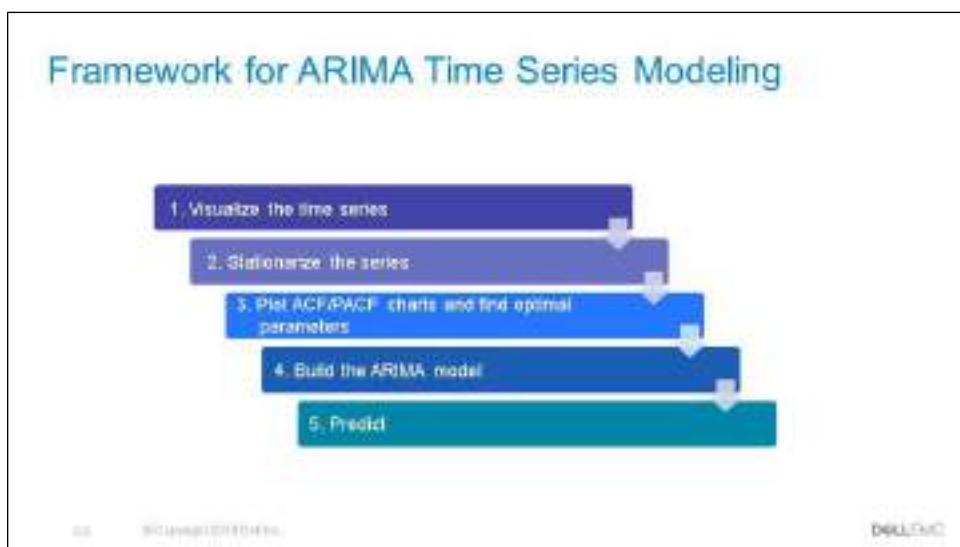


DATA SOURCE: DELL INC.

Here, you see a simple model for the time series with the trend, seasonality, and a random fluctuation. There is sometimes a low frequency cyclic term, as well, but you can ignore that, for the sake of simplicity.

Examples of trend and seasonality are also detailed in the slide.

## Framework for ARIMA Time Series Modeling



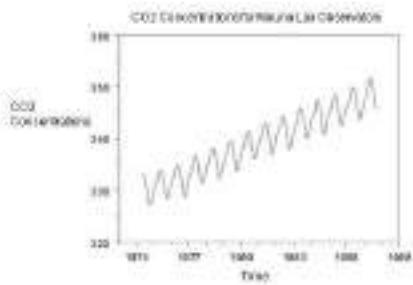
Here, you see the steps involved in Time Series Modeling.

- Visualize the time series
- Stationarize the series
- Plot ACF/PACF charts and find optimal parameters
- Build the ARIMA model
- Predict

## Step 1—visualizing time series

### Step 1—visualizing time series

- Analyze the trend before building any time series model.
- Details of interest pertain to any patterns.
  - Trend
  - Seasonality
  - Random behavior
- This graph is an upward trending time series with seasonality every year.



102      800-645-0000 | dell.com/Analytics

Dell EMC

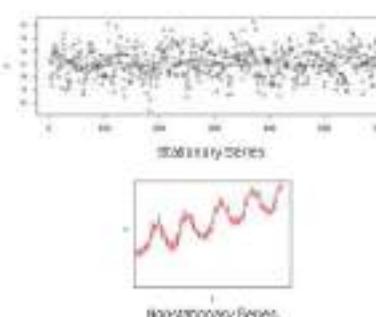
Step 1 in ARIMA time series modeling is visualizing the time series. Identifying patterns—trends, seasonal behavior, or any random behavior.

Here is an example of an upward trending time series with seasonality.

## Step 2—stationarize series

### Step 2—stationarize series

- Box-Jenkins methodology assumes that the random component is a stationary sequence:
  - Constant mean
  - Constant variance
  - Autocorrelation does not change over time.
    - » Constant correlation of a variable with itself at different times
- In practice, to obtain a stationary sequence, the data must be:
  - Detrended
  - Seasonally adjusted



stationary series

Nonstationary Series

DELL EMC

A stationary sequence is a random sequence in which the joint probability distribution does not vary over time. In other words, the mean, variance, and autocorrelations do not change in the sequence over time.

To render a sequence stationary, you must remove the effects of trend and seasonality. The ARIMA model, implemented with Box-Jenkins, uses the method of differencing to render the data stationary.

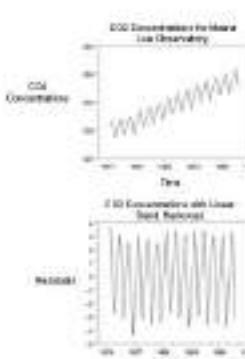
## Step 2—stationarize series—detrending

### Step 2—stationarize series—detrending

In this example, there is a linear trend, so fit a linear model:  
 $T_t = \alpha + \beta t$

The detrended series is then:  
 $\hat{Y}_t = Y_t - T_t$

Sometimes, you may have to fit a non-linear model:  
Quadratic  
Exponential



DATA SOURCE: DELL INC.

Trend in a time series is a slow, gradual change in some property of the series over the whole interval under investigation.

Detrending is a preprocessing step to prepare time series for analysis by methods that assume stationarity.

A simple linear trend can be removed by subtracting a least-squares-fit straight line. In the example shown, you fit a linear model and obtain the difference. The graph shown next is a detrended time series.

More complicated trends might require different procedures such as fitting a nonlinear model such as a quadratic or an exponential model.

Use a **Linear Trend Model** if the first differences are more or less constant [  $(y_2 - y_1) = (y_3 - y_2) = \dots = (y_n - y_{n-1})$  ]

Use a **Quadratic Trend Model** if the second differences are more or less constant.  
[  $(y_3 - y_2) - (y_2 - y_1) = \dots = (y_n - y_{n-1}) - (y_{n-1} - y_{n-2})$  ]

Use an **Exponential Trend Model** if the percentage differences are more or constant. [  $((y_2 - y_1)/y_1) \times 100\% = \dots ((y_n - y_{n-1})/y_{n-1}) \times 100\%$  ]

## Step 2—stationarize series—seasonal adjustment

Step 2—stationarize series—seasonal adjustment

Plotting the detrended series identifies seasonal patterns.

For CO<sub>2</sub> concentration, you can model the period as being a year, with variation at the month level.

Simple improved adjustment: take several years of data, calculate the average value for each month, and subtract that from  $y_t$ .

$$y_t^* = y_t - \bar{S}_t$$

140 © Copyright 2018 Dell Inc. DELL.COM

Unlike the trend and cyclical components, seasonal components, theoretically, happen with similar magnitude during the same time period each year.

The holiday sales spike is an example of seasonality. By removing the seasonal component, it is easier to focus on other components. The seasonal component of a series typically makes the interpretation of a series more difficult.

A simple adjustment for seasonality is accomplished by taking several years of data, calculating average value for each month, and subtracting them from the actual value.

## Step 3—plot ACF and PACF to identify optimal parameters

### Step 3—plot ACF and PACF to identify optimal parameters

- Auto Correlation Function (ACF)
  - Correlation of the values of the time series with itself
  - Autocorrelation “carries over”
  - Helps to determine the order,  $q$ , of a MA model
    - Where does ACF go to zero?
- Partial Auto Correlation Function (PACF)
  - An autocorrelation calculated after removing the linear dependence of the previous terms
  - Helps to determine the order,  $p$ , of an AR model
    - Where does PACF go to zero?

DATA SCIENCE & BIG DATA ANALYTICS

DELL EMC

A common assumption in many time series techniques is that the time series is stationary. A stationary process has the property that the mean, variance, and autocorrelation structure do not change over time.

An ACF plot provides an indication of the stationarity of the data. If the time series is not stationary, you can often transform it to stationarity with the simple technique of differencing. It should be noted that the autocorrelation carries over; if  $Y_t$  is correlated with  $Y_{t-1}$ , it is also correlated with  $Y_{t-2}$ , though to a lesser degree.

For a PACF, the partial autocorrelation at lag  $k$  is the autocorrelation between  $Y_t$  and  $Y_{t-k}$  that is not accounted for by lags 1 through  $k-1$ .

Look for the point on the plot where the partial autocorrelations for all higher lags are essentially zero.

You will look into ACF and PACF graphs in the next lab.

## Step 4—build model—ARMA (p, q)

### Step 4—build model—ARMA (p, q)

$$Y_t = \boxed{\delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p}} \\ \boxed{+ \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q}}$$

- $Y_t$  is detrended and seasonally adjusted.
- Combination of two process models
  - Autoregressive:  $Y_t$  is a linear combination of its last  $p$  values.
  - Moving average:  $Y_t$  is a constant value plus the effects of a damped white noise process over the last  $q$  time values—lags.

100

DATA SCIENCE

DATA SCIENCE

Autoregressive (AR) models can be coupled with moving average (MA) models to form a general and useful class of time series models called *Autoregressive Moving Average (ARMA)* models. The ARMA model is the simplest Box-Jenkins model.

AR model predicts  $Y_t$  as a linear combination of its last  $p$  values. An autoregressive model is simply a linear regression of the current value of the series on one or more prior values of the same series. Several options are available for analyzing autoregressive models, including standard linear least squares techniques. They also have a straightforward interpretation.

The time series  $Y_t$  is called an autoregressive process of order  $p$  and is denoted as AR( $p$ ) process.

A moving average (MA) model adds to  $Y_t$  the effects of a damped white noise process over the last  $q$  steps. The simple moving average is one of the most basic of the forecasting methods. Moving backwards in time, minus 1, minus 2, minus 3, and so forth, until you have  $n$  data points, divide the sum of those points by the number of data points,  $n$ , and that gives you the forecast for the next period. So, it is called a single moving average or simple moving average. The forecast is simply a constant value that projects the next time period. "n" is also the order of the moving averages.

A moving average is similar to a random walk, or Brownian motion.

## Step 4—build model—ARIMA (p, d, q)

### Step 4—build model—ARIMA (p, d, q)

- A stochastic modeling approach that can be used to calculate the probability of future value lying between two specified limits
- ARIMA adds a differencing term, d, to the ARMA model
  - Autoregressive Integrated Moving Average
  - Includes the detrending as part of the model:
    - Linear trend can be removed by  $d=1$
    - Quadratic trend can be removed by  $d=2$
    - And so on, for higher-order trends
- The general nonseasonal model is known as ARIMA (p, d, q):
  - p is the number of autoregressive terms.
  - d is the number of differences.
  - q is the number of moving average terms.

DATA SCIENCE

DELL EMC

ARMA models can be used when the series is **weakly stationary**; in other words, the series has a constant variance around a constant mean. This class of models can be extended to nonstationary series by allowing the differencing of the data series. These models are called *Autoregressive Integrated Moving Average* (ARIMA) models. There are a large variety of ARIMA models.

ARIMA—difference the  $Y_t$  d times to induce stationarity. d is usually 1 or 2. "I" stands for integrated; the outputs of the model are summed up—or, integrated—to recover  $Y_t$ .

The general ARIMA (p, d, q) model gives a tremendous variety of patterns in the ACF and PACF, so it is not practical to state rules for identifying general ARIMA models. In practice, it is seldom necessary to deal with values of p, d, or q other than 0, 1, or 2. It is remarkable that such a small range of values for p, d, or q can cover such a large range of practical forecasting situations.

## Step 4—build model—model selection

### Step 4—build model—model selection

- + Based on the data, the Data Scientist selects  $p$ ,  $d$ , and  $q$ :
  - An "art form" that requires domain knowledge, modeling experience, and a few iterations
  - Use a simple model when possible:
    - o AR model ( $q = 0$ )
    - o MA model ( $p = 0$ )
- + Multiple models must be built and compared, using:
  - ACF and PACF
  - $p$ -values for the model parameter estimates
  - Information criteria metrics:
    - o Imposes a penalty based on number of model parameters
    - o Example: Akaike Information Criterion (AIC)



© 2018 Dell Inc. All rights reserved.

Dell EMC

Identification of the most appropriate model is the most important part of the process, where it becomes as much "art" as "science."

The first step is to determine if the time series is stationary. This determination can be done with a correlogram, plots of the ACF and PACF. If the time series is not stationary, it must be first-differenced; it may need to be differenced again to induce stationarity.

The next stage is to determine the  $p$  and  $q$  in the ARIMA ( $p, d, q$ ) model. The  $d$  refers to how many times the data must be differenced to produce a stationary series.

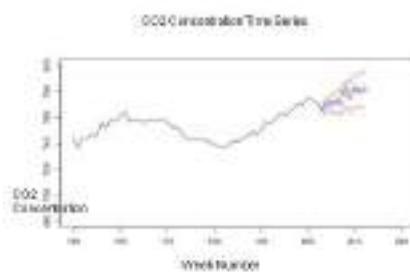
In the diagnostic stage, you assess the model's adequacy by checking whether the model assumptions are satisfied. If the model is inadequate, this stage provides some information for you to reidentify the model. You also perform checking normality, constant variance, and independence assumption among residuals.

## Step 5—predict

### Step 5—predict

After you have the final ARIMA model, you can use this model to make future predictions with the time series.

- Here is an example output of prediction using ARIMA model. You are predicting CO<sub>2</sub> concentrations for the next four weeks, based on historical time series. The blue dots to the right of the graph indicate the prediction, and the two red lines indicate the prediction interval.



## Time Series Analysis—reasons to choose (+) and cautions (-)

| Time Series Analysis—reasons to choose (+) and cautions (-)                                      |   |
|--|---|
| REASONS TO CHOOSE (+)  | CAUTIONS (-)  |
| Minimal data collection<br>Correlate the series itself.<br>It is not necessary to input drivers. | Stochastic signal drivers: prediction is based only on past performance<br>No explanatory value<br>Can handle "what-if" scenarios<br>Correlation test |
| Designed to handle time lags:<br>auto-correlation of lagged time series                          | It is an "informal" technique; appropriate parameters   |
| Accounts for trend and seasonality   | Only suitable for short-term predictions  |

The reasons to choose (+) and cautions (-) of Time Series Analysis are listed.

Time Series Analysis is not a common "tool" in a Data Scientist's toolkit. Though the models require minimal data collection and handle the inherent auto correlations of lagged time series, it does not produce meaningful drivers for the prediction.

The selection of (p, d, q) appropriately is not straightforward. A complete understanding of the domain knowledge and detailed analysis of trend and seasonality may be needed. Further, this method is suitable for short-term predictions only.

## Check your knowledge

### Check your knowledge

1. What is a time series, and what are the key components of a time series?
2. How do you detrend time series data?
3. What makes data stationary?
4. How is seasonality removed from the data?
5. What are the modeling parameters in ARIMA?
6. How do you use ACF and PACF to determine the stationarity of time series data?



## Check your knowledge



### Discussion

## Question / Discussion Topic:

1. What is a time series, and what are the key components of a time series?
2. How do you detrend time series data?
3. What makes data stationary?
4. How is seasonality removed from the data?
5. What are the modeling parameters in ARIMA?
6. How do you use ACF and PACF to determine the stationarity of time series data?

## Discussion Notes:

## Time series analysis—summary

### Time series analysis—summary

During this lesson, the following topics were covered:

- Time series analysis and its applications in forecasting
- ARMA and ARIMA Models
- Implementing the Box-Jenkins methodology using R
- Reasons to choose (+) and cautions (-) with time series analysis



14      80 minutes estimated

Dell EMC

This lesson covered these topics. Take a moment to review them.

## Module summary

| Module summary   |   |
|--|---|
| Key concepts introduced in this module:                    | Variables, continuous vs discrete   |
| Algorithms and statistical foundations:                    | Classification, supervised:<br>Hierarchical clustering<br>Association rules |
| Key applications:  | Prediction:<br>Linear<br>Logistic   |
| Classification and validation of the model:                | Classification, supervised:<br>Naive Bayes and logistic<br>Decision trees   |
| Requirements (both (+) and (minus)) of the model:          | Time series analysis  |
| Final tip: avoiding common mistakes in data distributions: | Test statistics   |

A summary of the key topics presented in this module is listed here.

## Advanced analytics—technology and tools

### Introduction



### Advanced analytics—technology and tools

Upon completing this module, you should be able to:

- ✓ Perform analytics on unstructured data using the MapReduce Programming paradigm.
- ✓ Use Hadoop, Hive, and Pig for unstructured data analytics.
- ✓ Effectively use advanced SQL functions and the Greenplum extensions for in-database analytics.
- ✓ Use MADlib to solve analytics problems in-database.

The objectives for this module are shown here.

Upon completing this module, you should be able to:

- Perform analytics on unstructured data using the MapReduce Programming paradigm.
- Use Hadoop, Hive, and Pig for unstructured data analytics.
- Effectively use advanced SQL functions and the Greenplum extensions for in-database analytics.
- Use MADlib to solve analytics problems in-database.

## Lesson: Introduction to advanced analytics— technology and tools

Introduction



## Challenges with Big Data beyond analytics

**Challenges with Big Data beyond analytics**

- Infrastructure
  - Storage
  - Backups/restores
  - Compute
  - Network
- Architectural complexities
- Security
  - Data governance
  - Regulatory compliance
- Data quality
- Ethics

The diagram consists of two blue circles. The left circle contains the text "Technology and tools" and has a blue arrow pointing clockwise to the right circle. The right circle contains the text "Business processes" and has a blue arrow pointing clockwise back to the left circle.

DATA SOURCE: DELL INC.

The emerging data ecosystem demonstrates a new economy that is emerging around data—a situation in which data has intrinsic value. The real-time analysis of the data gathered, coupled with advanced analytical methods, demands new analytical architectures for storing and analyzing these kinds of data.

It is important to recognize some of the other Big Data challenges beyond analytics.

In the **infrastructure** space:

- How will all of the data be stored and backed up?
- How much computing power will be required to process the data?
- How will the network handle the possible movement of data from the source systems to the analytic sandboxes?
- How will the infrastructure scale over time?
- Can computation and storage scale independently?

### Architectural complexities

The vast proliferation of technologies in this competitive market means there is no single go-to solution when you begin to build the Big Data architecture. An examination of the Big Data ecosystem shows that there are various technologies that exist to harness their data.

## Lesson: Introduction to advanced analytics—technology and tools

Infrastructure serves to process, store, and enable data analysis. The rise of unstructured data in particular meant that data capture had to move beyond merely rows and tables. So, new infrastructural technologies emerged, capable of wrangling a vast variety of data and making it possible to run applications on systems with thousands of nodes, potentially involving thousands of terabytes of data.

### **Analytics**

Although infrastructural technologies incorporate data analysis, there are specific technologies that are designed specifically with analytical capabilities. Analytical, Business Intelligence, Visualization and Machine Learning Platform

### **Applications**

Big Data businesses leverage applications to offer users optimized analytic insights.

### **Security**

In addition to ensuring that hackers cannot access the vast repositories of data:

- How to ensure that only the proper users can access the appropriate data?
- How to comply with privacy laws and regulations?

Furthermore, it is important to monitor and address the quality of the data and to ensure that the data is used in an ethical manner.

Obviously, technology and tools play an important part in addressing these challenges, but it is worth noting that business processes are tightly coupled with how the tools are used.

## What is Apache Hadoop?

### What is Apache Hadoop?

- Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.<sup>2</sup>
- The Apache Hadoop software library is a framework that allows for the distributed processing of large datasets across clusters of computers using simple programming models.<sup>3</sup>
- This library enables us to use parallel processing capability to handle huge volumes of data using flexible infrastructure.
- To summarize, Hadoop offers:
  - A scalable, flexible, and reliable distributed computing Big Data framework for a cluster of systems.
  - Storage capacity.
  - Local computing power—applying commodity hardware.
- Hadoop is not:
  - A database.
  - Simply a data warehouse tool. But it can be used as one.
- Hadoop is written in Java™.

343

© Copyright 2018 Dell Inc.

DELL.COM

\* Source: [hadoop.apache.org](http://hadoop.apache.org)

Unfortunately, people may use the word “Hadoop” to mean multiple things. They may use it to describe massive unstructured data storage using commodity hardware—although commodity does not mean inexpensive—or a data warehouse tool.

On the other hand, they may be referring to the Java classes provided by Hadoop that support HDFS file types or provide MapReduce job management.

But Hadoop is an ecosystem of open-source projects that provides framework to deal with Big Data and help us handle the three V’s—volume, velocity, and variety. Hadoop enables us to use parallel processing capability to handle huge volumes of data using flexible commercial-grade infrastructure.

Hadoop offers:

- A scalable, flexible, and reliable distributed computing Big Data framework for a cluster of systems.
- Storage capacity.
- Local computing power—applying commodity hardware.

## Lesson: Introduction to advanced analytics—technology and tools

Remember that Hadoop is not:

- Simply another data warehouse tool. But it is used as one.
- A database.

Hadoop is written in Java™.

## Four main components of Apache Hadoop

### Four main components of Apache Hadoop

The diagram illustrates the four main components of Apache Hadoop as stacked horizontal bars. From top to bottom, they are: MapReduce—data processing, YARN—resource management, HDFS—storage redundant, and Hadoop Common.

- MapReduce
- Yet Another Resource Negotiator (YARN™)
- Hadoop Distributed File System (HDFS™)
- Hadoop Common module is a Hadoop Base API —a jar file—for all Hadoop components. All other components work on top of this module.

The components are explained in detail in the next few slides.

144      © Copyright 2018 Dell Inc.      DELL.COM

**Apache Hadoop** is a powerful tool for Big Data. Hadoop consists of four main components: HDFS, MapReduce, YARN, and Core. Apart from these Hadoop components, there are some other Hadoop ecosystem components, also, that play an important role to boost Hadoop functionalities.

Hadoop comes with MapReduce. Other data processing modules can be used in addition to MapReduce (e.g., Pig, Spark, Hive, and others).

## Hadoop Distributed File System

### Hadoop Distributed File System

- Distributed file system designed to run on commodity hardware for storing large files of data with streaming data access patterns
- Highly fault tolerant
- Default storage for the Hadoop cluster
- File system namespace
- Data/File on HDFS is stored in chunks (128 MB default) called blocks

34

© Copyright 2018 Dell Inc.

DELL.COM

HDFS is a **distributed file system** that provides high-performance access to **data** across Hadoop clusters. As is true for other Hadoop-related technologies, HDFS has become a key tool for managing pools of **Big Data** and supporting **Big Data analytics** applications.

### Fault tolerant

Hardware failure is the norm rather than the exception. An HDFS instance may consist of hundreds or thousands of server machines, each storing part of the file system's data. The fact that there are a huge number of components and that each component has a nontrivial probability of failure means that some component of HDFS is always nonfunctional. Therefore, detection of faults and quick, automatic recovery from them is a core architectural goal of HDFS.

HDFS is the default storage for the Hadoop cluster.

### File system namespace

HDFS supports a traditional hierarchical file organization. A user or an application can create directories and store files inside these directories. The file system namespace hierarchy is similar to most other existing file systems; one can create and remove files, move a file from one directory to another, or rename a file. HDFS does not yet implement user quotas. HDFS does not support hard links or soft links. However, the HDFS architecture does not preclude implementing these features.

## Lesson: Introduction to advanced analytics—technology and tools

Data/file on HDFS is stored in chunks (128 MB default) called blocks.

Reference: [hadoop.apache.org/docs/r2.6.0/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html](https://hadoop.apache.org/docs/r2.6.0/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html)

## Hadoop Distributed File System—assumption/goals

### Hadoop Distributed File System—assumption/goals

- Hardware failure is a norm rather than an exception
- Streaming data access
- Large dataset processing
- Simple coherency model
- Moving computation is cheaper than moving data
- Portability across heterogeneous hardware and software platforms

30

DATA SCIENCE

DELL EMC

### Hardware failure

Hardware failure is the norm rather than the exception. An HDFS instance may consist of hundreds or thousands of server machines, each storing part of the file system's data. The fact that there are a huge number of components and that each component has a nontrivial probability of failure means that some component of HDFS is always nonfunctional. Therefore, detection of faults and quick, automatic recovery from them is a core architectural goal of HDFS.

### Streaming data access

Applications that run on HDFS need streaming access to their datasets. They are not general-purpose applications that typically run on general purpose file systems. HDFS is designed more for batch processing rather than interactive use by users. The emphasis is on high throughput of data access rather than low latency of data access. POSIX imposes many hard requirements that are not needed for applications that are targeted for HDFS. POSIX semantics in a few key areas has been traded to increase data throughput rates.

### Large datasets

Applications that run on HDFS have large datasets. A typical file in HDFS is gigabytes to terabytes in size. Thus, HDFS is tuned to support large files. It should provide high aggregate data bandwidth and scale to hundreds of nodes in a single cluster. It should support tens of millions of files in a single instance.

### **Simple coherency model**

HDFS applications need a write-once-read-many access model for files. After a file is created, written, and closed, it need not be changed. This assumption simplifies data coherency issues and enables high throughput data access. A MapReduce application or a web crawler application fits perfectly with this model. There is a plan to support appending-writes to files in the future.

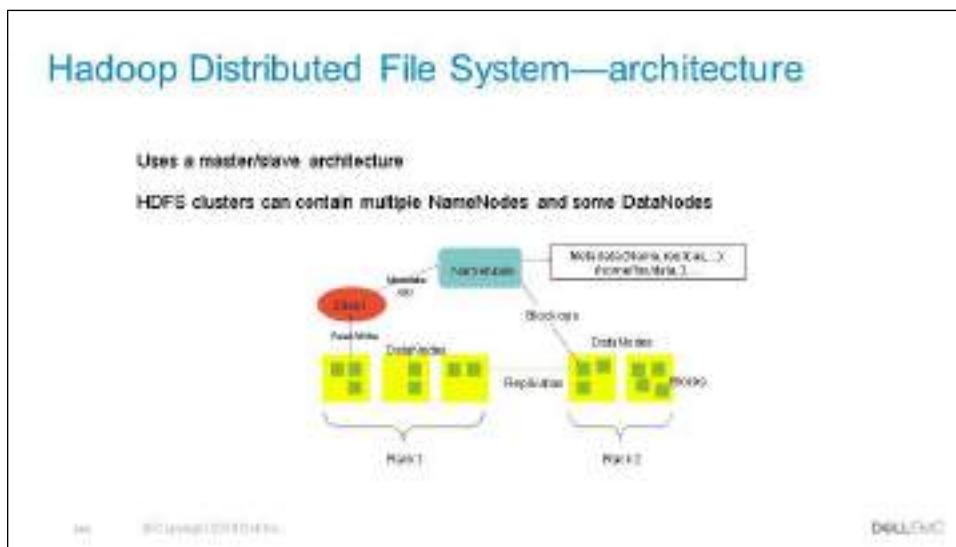
### **Moving computation is cheaper than moving data**

A computation requested by an application is much more efficient if it is executed near the data it operates on. This is especially true when the size of the dataset is huge. This proximity minimizes network congestion and increases the overall throughput of the system. The assumption is that it is often better to migrate the computation closer to where the data is located rather than moving the data to where the application is running. HDFS provides interfaces for applications to move themselves closer to where the data is located.

### **Portability across heterogeneous hardware and software platforms**

HDFS has been designed to be easily portable from one platform to another. This facilitates widespread adoption of HDFS as a platform of choice for a large set of applications.

## Hadoop Distributed File System—architecture



HDFS has a master/slave architecture. An HDFS cluster consists of a multiple NameNode that manages the file system namespace and regulates access to files by clients. Further, some DataNodes, usually one per node in the cluster, manage storage attached to the nodes that they run on.

The DataNodes are used as common storage for blocks by all the NameNodes. Each DataNode registers with all the NameNodes in the cluster. DataNodes send periodic heartbeats and block reports. They also handle commands from the NameNodes.

HDFS exposes a file system namespace and allows user data to be stored in files. Internally, a file is split into one or more blocks, and these blocks are stored in a set of DataNodes. The NameNode executes file system namespace operations such as opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes. The DataNodes are responsible for serving read and write requests from the file system's clients. The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode.

### References:

- [codrspace.com/benjibc/hdfs-overview](http://codrspace.com/benjibc/hdfs-overview)
- [hadoop.apache.org/docs/r1.2.1/hdfs\\_design.html](http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html)
- [hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/Federation.html](http://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/Federation.html)

## NameNode

### NameNode

- A master server that manages the file system namespace and regulates access to clients.
- NameNode maps the entire file system and metadata structure—such as permissions, modification timestamp, and so on—into memory.
- It executes file system namespace operations such as opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes.
- Information is stored persistently on the disk in the form of two files: namespace image and edit log.
  - Namespace image file contains the Inodes and the list of blocks which define the metadata.
  - Edit log contains any modifications that have been performed on the content of the image file.

44

© Copyright 2018 Dell Inc.

DELL EMC

NameNode is the centerpiece of an HDFS file system. It keeps the directory tree of all files in the file system, and tracks where across the cluster the file data is kept. It does not store the data of these files itself.

Client applications talk to the NameNode whenever they want to locate a file, or when they want to add, copy, move, or delete a file. The NameNode responds to the successful requests by returning a list of relevant **DataNode** servers where the data lives.

#### Tasks of HDFS NameNode:

- Manage the file system namespace.
- Regulate the client's access to files.
- Execute the file system execution such as naming, closing, and opening files and directories.

Class Inode holds file metadata including the type—regular file or directory—and the list of blocks that are pointers to the data.

Information is stored persistently on the disk in the form of two files: namespace image and edit log.

Namespace image file (fsimage) contains the Inodes and the list of blocks which

## Lesson: Introduction to advanced analytics—technology and tools

define the metadata.

Edit log contains any modifications that have been performed on the content of the image file.

## DataNodes

### DataNodes

- A file is split into one or more blocks, and these blocks are stored in multiple DataNodes.
- DataNodes are responsible for serving read and write requests from the file system's clients.
- The DataNodes perform block creation, deletion, and replication upon instruction from the NameNode.
- Reports back to NameNode with the list of stored blocks.
- Bringing computation to data is often more efficient than the reverse.
- DataNodes perform most CPU-intensive and I/O-intensive jobs.

101

DATA SCIENCE & ANALYTICS

DELL EMC

DataNode is also known as *Slave*. HDFS DataNode is responsible for storing actual data in HDFS. DataNode performs read and write operations as per the request of the clients.

Tasks of HDFS DataNode:

- DataNode performs operations such as block replica creation, deletion, and replication according to the instruction of NameNode.
- DataNode manages data storage of the system.
- DataNodes perform CPU-intensive jobs such as semantic and language analysis, statistics, and machine learning tasks, as well as I/O-intensive jobs including clustering, data import, data export, search, decompression, and indexing. And, they report back to NameNode with the list of blocks they are storing.

Bringing computation to data is often more efficient than the reverse. The resources on a data node are typically made available to run computing processes.

## Block replication

### Block replication

Data is replicated more than once in a Hadoop cluster to fault tolerance and availability. Every block of data is replicated on more than one node in, even if a node fails, the data is available on another node.

The replication factor is the number of times a block is replicated. The default is 3 for HDFS, which means every block is replicated three times on three different nodes—see Example below. (D0 stands for data block.)

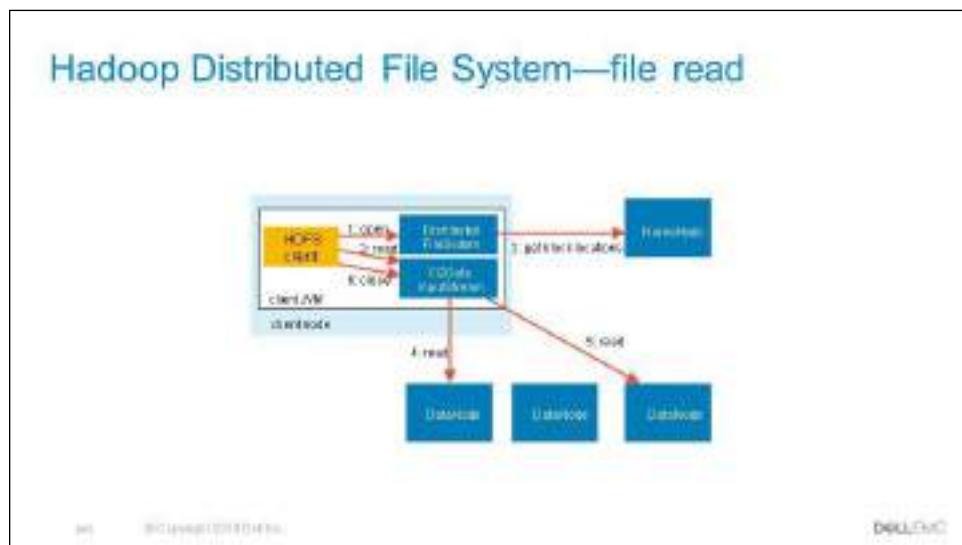
D0 D1 D2 D3  
Node 1 Node 2 Node 3 Node 4  
HDFS

HDFS is designed to reliably store very large files across machines in a large cluster. It stores each file as a sequence of blocks; all blocks in a file except the last block are the same size. The blocks of a file are replicated for fault tolerance. The block size and replication factor are configurable per file.

An application can specify the number of replicas of a file. The replication factor can be specified at file creation time and can be changed later. Files in HDFS are write-once and have no more than one writer at any time.

The default replication factor for HDFS is 3.

## Hadoop Distributed File System—file read



Source: [coderspace.com/benjibc/hdfs-overview](https://coderspace.com/benjibc/hdfs-overview)

### Step 1:

The client opens the file it aims to read by calling `open` on the file system object, which for HDFS is an instance of `DistributedFileSystem`.

### Step 2:

`DistributedFileSystem` calls the `NameNode` to determine the locations of the blocks for the first few blocks in the file. For each block, the `NameNode` returns the addresses of the `DataNodes` that have a copy of that block, and `DataNodes` are sorted according to their proximity to the client. `DistributedFileSystem` returns a `DataInputStream` to the client for it to read data from.

### Step 3:

Client calls `read` on the stream. `FS Data InputStream`, which has stored the `DataNode`, addresses and then connects to the closest `DataNode` for the first block in the file.

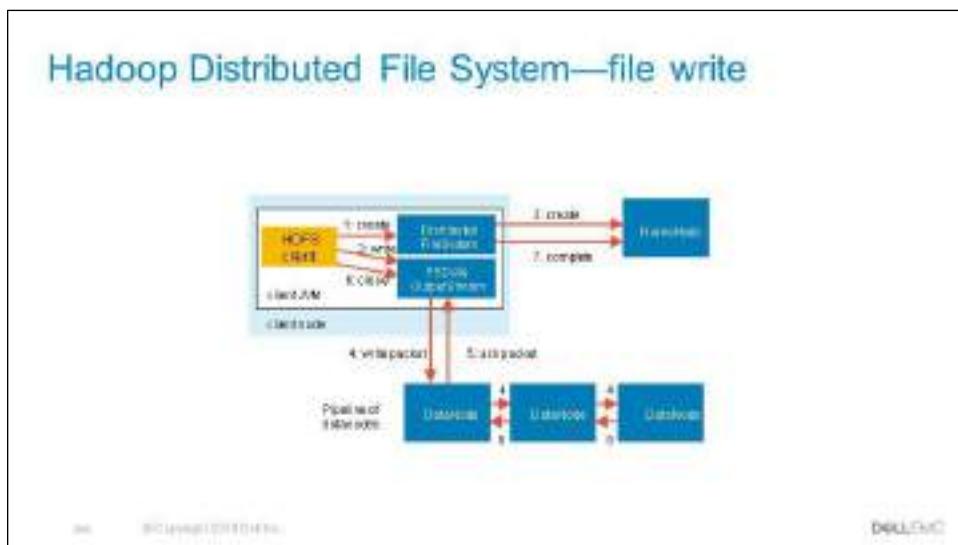
### Steps 4 and 5:

Data is streamed from the `DataNode` back to the client, which calls `read` repeatedly on the stream. When the end of the block is reached, `FSInputStream` closes the connection to the `DataNode` and then finds the best `DataNode` for the next block.

## Lesson: Introduction to advanced analytics—technology and tools

If the FSInputStream encounters an error while communicating with a DataNode, it will try the next closest one for that block. It will also remember DataNodes that have failed, so that it does not needlessly retry them for later blocks. The FSInputStream also verifies checksums for the data transferred to it from the DataNode. If a corrupted block is found, it is reported to the NameNode before the DFSInputStream attempts to read a replica of the block from another DataNode.

## Hadoop Distributed File System—file write



**Source:** [codrspace.com/benjibc/hdfs-overview](http://codrspace.com/benjibc/hdfs-overview)

### Step 1:

Before the client starts writing data to HDFS, it grabs an instance of an object of Distributed File System (HDFS).

### Step 2:

The `DistributedFileSystem` object calls `NameNode` to create a new file in file system namespace with no blocks associated to it.

`NameNode` process performs various checks, such as:

- If client has required permissions to create a file.
- If file existed earlier—it should not have. If it did, the `NameNode` throws an IO exception to client.

### Step 3:

After the file is registered with the `NameNode`, then the client gets an object—that is, `FSDataOutputStream`—which, in turn, embeds the `DFSSoutputStream` object for the client to start writing data to. The `DFSSoutputStream` handles communication with the `DataNodes` and `NameNode`.

### Steps 4 and 5:

## Lesson: Introduction to advanced analytics—technology and tools

As client writes data, the DFSOutputStream splits it into packets and writes it to its internal queue—that is, data queue—and also maintains an acknowledgement queue.

The data queue is then consumed by a data streamer process that is responsible for asking NameNode to allocate new blocks by picking a list of suitable DataNodes to store the replicas.

The list of DataNodes forms a pipeline and assumes a replication factor of three, so there will be three nodes in the pipeline. The data streamer streams the packets to the first DataNode in the pipeline, which then stores the packet and forwards it to the second DataNode in the pipeline. Similarly, the second node stores the packet and forwards it to the next DataNode or last DataNode in the pipeline.

After each DataNode in the pipeline acknowledges the packet, the packet is removed from the acknowledgement queue.

## Hadoop Distributed File System—not ideal in the following situations

### Hadoop Distributed File System—not ideal in the following situations

- Low-latency reads
  - High throughput rather than low latency for small chunks of data
- Large number of small files
  - Better for large files but not for millions of small files
- Multiple writes
  - Single write per file
  - Writes only at the end of the file

181

DATA SCIENCE WITH HDFS

DELL EMC

HDFS is not for:

- Low-latency reads
  - High throughput rather than low latency for small chunks of data
- Large number of small files
  - Better for large files but not for millions of small files
- Multiple writes
  - Single write per file
  - Writes only at the end of the file

## Introduction to MapReduce

The slide has a light gray header bar with the title 'Introduction to MapReduce'. Below it is a white rectangular area containing a blue rounded rectangle. Inside the blue rectangle is a quote from Grace Hopper:

**"In pioneer days, they used oxen for heavy pulling. When one ox couldn't budge a log, they didn't try to grow a larger ox..."**

**We shouldn't be trying to grow bigger computers, but to add more systems of computers."**

- Grace Hopper

Below the quote, a small note states: "The Hadoop ecosystem with the HDFS and MapReduce paradigm helps you add more oxen". At the bottom of the slide, there are small text elements: '14 BIG DATA WITH HADOOP', 'DELL EMC', and 'DATA SCIENCE'.

You see from Grace Hopper's quote that the fundamental paradigm of MapReduce is the **reduction in time to complete a given task by breaking it down into stages and then executing those stages in parallel**.

Such an activity is sometimes called the master/slave or master/worker pattern, and it has been known for a while.

The definition of Big Data asserts that the data is simply too large to handle by conventional means. The usual example is when an RDBMS is initially used to store that data. As performance needs increase, organizations purchase more powerful hardware and then more systems to share the data retrieval and processing. And, yet, the data keeps on growing.

What can be done?

## What MapReduce is

### What MapReduce is

- A software paradigm for writing applications that process vast amounts of data, multi-terabyte datasets, in-parallel on large clusters—thousands of nodes—of commodity hardware in a reliable, fault-tolerant manner
- Java-based programming paradigm
- A combination of the Map and Reduce models that can be applied to wide variety of business cases
- Handles scheduling and fault tolerance

## When to use MapReduce

### When to use MapReduce

- Problems that are “embarrassingly parallel”
- Examples
  - Word count
  - Reverse index
  - Tf-idf
  - Distributed grep and distributed object recognition—“Where’s Waldo?”
  - Distributed associative aggregation—marginalization, sum, mean if you track both numerator and denominator; min or max; count

10

DATA SCIENCE

DELL EMC

For which kinds of problems is MapReduce most suited? Simply, problems that are “embarrassingly parallel.” In these situations, the problem can be decomposed into tasks that can be broken into “chunks” that can be distributed, processed, and recombined without communicating with each other.

Examples of such problems are listed here. You have met some of these functions before: reverse index, tf-idf, distributed grep—pattern matching.

**Marginalization occurs when you use aggregate statistics over certain variables as a way of reducing the number of variables in an analysis by focusing on the aggregation of other variables.**

Consider this example. Assume that you have a table where each object has the attributes “state”, “climate”—such as wet, dry, temperate—and “% of cats in the US”, or *pct\_cats*. This table provides us with the number of cats broken out by climate and state. If you sum the *pct\_cats* across each state, then you are calculating the marginal probability of cats in each state. If you sum *pct\_cats* across each value of climate, you are calculating the marginal probability of cats as a function of climate.

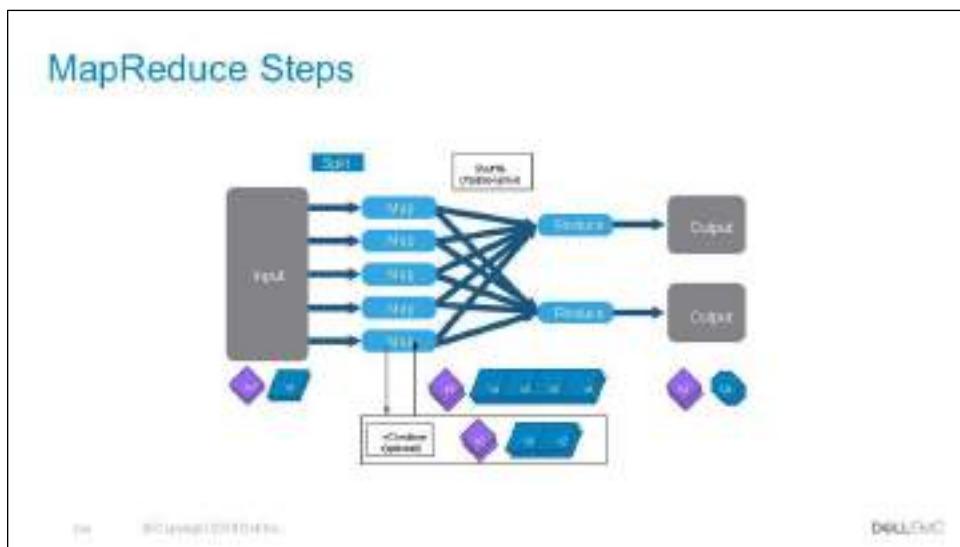
In statistical terms, the table gives you  $P(\text{cats} \mid \text{state}, \text{climate})$ . Marginalization aggregates this table up to either  $P(\text{cats} \mid \text{state})$ , or  $P(\text{cats} \mid \text{climate})$ , depending. For the SQL inclined, the statement

## Lesson: Introduction to advanced analytics—technology and tools

*SELECT SUM(pct\_cats) from table GROUP BY state*

would perform this action for the values of state.

## MapReduce Steps



The MapReduce starts with an input of raw data. When the raw data is feed to the system via an input reader, the first set of key-value pairs ( $K_1, V_1$ ) is generated. An example of such ( $K_1, V_1$ ) pair would be (line\_number, line\_text).

The next step is to split the input for different mapper nodes to process. The mappers then produce the second key-value pair or ( $K_2, V_2$ ) which is the result of business logic map function. An example of a map function output for word counting is a list of (word, count). It is important to note that this output will not be an aggregated output for each key like ('Hamlet', 5). It would be a list of ( $K_2, V_2$ ) containing five ('Hamlet', 1)s. Sometimes a combiner function is provided to aggregate (reduce) mapper output, grouped by keys ( $K_2$ ). This would then provide an output similar to ( $K_2, V_2'$ ) or ( $K_2, \text{list}(V_2)$ ) depending on the context. For instance, the combiner would produce ('Hamlet', 5) or ('Hamlet', <1,1,1,1,1>).

The next step is shuffling (or partitioning) and sorting so that all ( $K_2, V_2$ s) for the same key, be provided to a single Reducer. This step may not be used in some use-cases.

In the last step, the reducer provides the output, based on the ( $K_2, V_2$ s) of **all** the mapper nodes and produces and output varying based on the business logic (usually saving the files to an output).

The final output ( $K_3, V_3$ ) is usually loaded back into the storage.

## MapReduce paradigm

### MapReduce paradigm

- Data processing system with two phases—Map and Reduce
- Map
  - Performs a map function on key input key-value pairs to generate intermediate key-value pairs
- Reduce
  - Performs a reduce function on intermediate key-value groups to generate output key-value pairs
- Process
  - Input: a set of key-value pairs.
  - User supplies two functions
    - Map  $(K_1, V_1) \rightarrow \text{list}(K_2, V_2)$
    - Reduce  $(K_2, \text{list}(V_2)) \rightarrow (K_3, V_3)$
  - $(K_2, V_2)$  is an intermediate key-value pair
  - Intermediate key-value groups are created by sorting map output which is input to combiner or reduce function.

— BY DELL INC.

The first phase of any MapReduce job is to, well, map.

In the mapping phase, raw data is transformed into a set of <key, value> pairs.

For example, the key might be a line number and the value a text string. The map function will transform this input into a series of output pairs: in this case, a record containing each unique word in the input—the key—and a count of its occurrences, known as the value. So, for example, the input string “to be or not to be” would result in a series of output records such as [to 2; be 2; or 1; not 1]—assume that the “;” character represents a new line.

Well, you have finished mapping, and now what?

The next task is to run the Reduce step on the input from the Mapper.

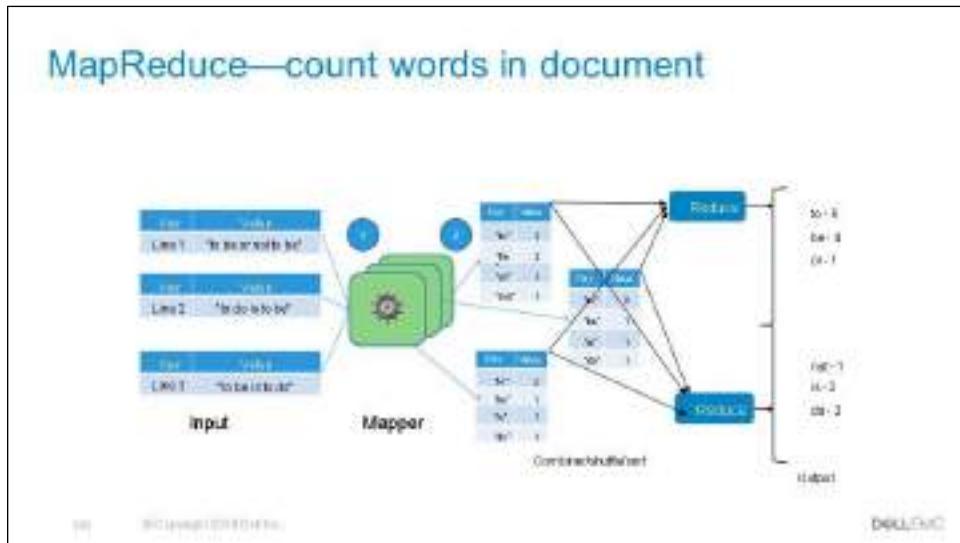
Sometimes, all the work is completed in the Mapper, and you are finished.

Otherwise, you use the Reducer to combine the input from the Mapper. Maybe you count it, print it, load it into a database, or save it in a file and load it into R for more analysis.

The MapReduce framework supports a combiner function, as well. This function would take Mapper output and produce a second output stream for the Reducer component.

For example, you could imagine a Combiner function that would take output from multiple mappers and further recombine it before handing it off to the Reducer job. **Combiners work well when Mapper output is voluminous: adding a Combiner step may decrease network traffic, resulting in better performance.** Combiner usage also reduces I/O as they get applied after merge sorts of spill files (done in memory) so the resultant disk I/O is reduced as well.

## MapReduce—count words in document



## Another word count example

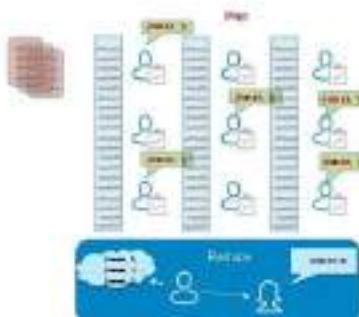
### Another word count example

This example is the “Hello, World” of MapReduce.

Distribute the text of millions of documents over hundreds of machines.

**Mappers** can be word-specific. They run through the stacks and shout “One!” every time they see the word “Hamlet”.

**Reducers** listen to all the Mappers and total the counts for each word.



© 2018 DELL INC. DELL.COM

The “word count” example is the “Hello, World” of the data analytics world. Here, you have millions of documents and hundreds of machines. You want to count the number of times the word “Hamlet” is displayed in these documents.

The key to understanding this problem is that the Mappers do not attempt to actually aggregate the count of the number of times the word “Hamlet” is displayed in a single document. Instead, they simply output a key-value pair consisting of <Hamlet,1>.

The combiner at each node, combines the output of mappers to create an aggregate key-value pair at each node (<Hamlet, 2> for the first, and the last mapper is the output of the combiner)

It is up to the reducers to aggregate the final results and output the single key-value pair <Hamlet, nnn>.

## Where MapReduce is used—some examples

### Where MapReduce is used—some examples

- Index building in search engines
- Article clustering for news
- Statistical machine translation
- Data mining
- Spam detection
- Ad optimization

100

DATA SCIENCE

DELL EMC

MapReduce is used on a regular basis for many companies dealing with large volume of data.

For example, Google is using MapReduce for building the index for its search engine. Article clustering and machine translation at Google are also heavily dependent on MapReduce.

Yahoo and Facebook are also using MapReduce for spam detection, indexing, and targeted advertisement.

## Example—social networking—eDiscovery

eDiscovery can involve processing a huge number of documents. Consider this example, where you want to process over one-half a million email messages, and you want to determine the social network in the company, where “social network” is a particular group of people who communicate:

**Date:** Thu, 4 May 2000 09:55:00 -0700 (PDT)

**From:** [walt.zimmerman@enron.com](mailto:walt.zimmerman@enron.com)

**To:** [michael.burke@enron.com](mailto:michael.burke@enron.com), [dana.gibbs@enron.com](mailto:dana.gibbs@enron.com), [lori.maddox@enron.com](mailto:lori.maddox@enron.com), [susan.ralph@enron.com](mailto:susan.ralph@enron.com)

**Subject:** Update on Steve Todoroff Prosecution—CONFIDENTIAL/SUBJECT TO ATTORNEY-CLIENT PRIVILEGE

**Cc:** steve.duffy@enron.com, stanley.horton@enron.com, ideageeter@velaw.com

*Almost one month ago, Special Agent Carl Wake of the FBI called me about the Steve Todoroff investigation. He indicated that the FBI had recently learned of the article about EOTT's NGL theft that appeared in the business section of the Houston Chronicle. Mr. Wake said it might be a matter the FBI would like to investigate. I told Mr. Wake that EOTT was currently working with the Harris County District Attorney on the prosecution of this matter, and I thanked him for the FBI's interest. He told me that the FBI might want to work with the Harris County District Attorney in investigating this matter, and he stated that there may be investigative information that the FBI can obtain more quickly than the Harris County District Attorney. Mr. Wake requested a copy of the materials we had provided to the*

## Lesson: Introduction to advanced analytics—technology and tools

*Harris County District Attorney.*

*In order to avoid damage to the good rapport we have established with Assistant District Attorney Bill Moore, I asked John DeGeeter to call Bill Moore and advise him of the contact that had been made by the FBI. Bill Moore agreed to call Carl Wake and work with Mr. Wake on his request for the materials provided by EOTT.*

*Carl Wake called me again yesterday. He has been working with Bill Moore. Mr. Wake stated it was too early to speculate as to what charges would be brought. He did say that our materials clearly indicated federal wire fraud and possibly mail fraud. He said that where there is wire fraud, there is usually money laundering.*

*The purpose of Mr. Wake's call yesterday was to inquire about the status of some interview summaries that John DeGeeter and I have prepared and collected at the request of Bill Moore. Mr. Wake requested that EOTT send a copy of the summaries to him when we sent the summaries to Bill Moore. Those summaries were sent out today.*

*I gathered from my calls with Carl Wake that the FBI is very interested in taking an active part in this investigation. In order to build on the relationship we have established with Bill Moore, we will continue to direct our inquiries about the investigation to Mr. Moore until he tells us to do otherwise.*

## Social triangle—first directed edge

**Social triangle—first directed edge**

**Mapper1**

- Maps two regular expression searches:
  - To Michael, Dan, Lori, Susan
  - From Walt
- Emits the outbound directed edge of the social graph:
  - <key, value> = <Walt, [Michael, Dan, Lori, Susan]>

**Reducer1**

- Gets the output from the mapper with different values:
  - <key, value> = <Walt, [Michael, Dan, Lori, Susan]>
  - <key, value> = <Walt, [Lori, Susan, Jeff, Ken]>
- Unites the values for the second directed edge:
  - <key, value> = <Walt, [Dan, Jeff, Ken, Lori, Michael, Susan]>

© Copyright 2018 Dell Inc.

DELL.COM

Build this job as a sequence of MapReduce passes through the data—you can assume that the modified data is written back into HDFS.

The output from Mapper1 is a key-value pair where the key is the sender of the message and the value is a list of the recipients.

The Reducer takes as input the list of sender/recipients and creates a single record for each sender with the aggregated list of all recipients. These are all the people to whom Walt sent email.

$\langle \text{Walt}, [\text{Michael, Dan, Lori, Susan}] \rangle + \langle \text{Walt}, [\text{Michael, Dan, Lori, Susan}] \rangle \rightarrow \langle \text{Walt}, [\text{Dan, Jeff, Ken, Lori, Michael, Susan}] \rangle$

## Social triangle—second directed edge

### Social triangle—second directed edge

**Mapper2**

- Reverses the previous map:
  - To: Michael, Dan, Lori, Susan
  - From: Walt
- Emits the inbound directed edge of the social graph:
  - <Key, Value> = <Susan, Walt>, <Lori, Walt>, <Dan, Walt>, and so on

**Reducer2**

- Gets the output from the mapper with different values:
  - <Key, Value> = <Susan, Walt>
  - <Key, Value> = <Susan, Jeff>
- Unions the values for the third directed edge:
  - <Key, Value> = <Susan, [Jeff, Ken, Walt]>

© 2014 Dell Inc.

Mapper2 does the opposite.

For each email recipient, it creates a key-value pair of the form <recipient/participant>.

Reducer2 is identical to Reducer1: for each recipient, it creates a single output record of the form <recipient/list of senders>. These represent people from whom the person received mail.

The fact that the second reducer is identical is a case of reused code. With MapReduce, simple data manipulation code can be easily reused.

Susan received email from [Jeff, Ken, Walt].

## Social triangle—third directed edge

### Social triangle—third directed edge

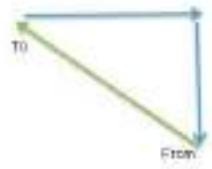
**Mapper3**

- Join [inbound] and [outbound] lists by key:
  - Walt, [Jeff, Ken, Lori, Susan], [Jeff, Lori, Stanley]
- Emits <Person, Person> pair with level of association:
  - <Key, Value> = <Walt:Jeff, reciprocal>; <Walt:Stanley, directed>, and so on

**Reducer3**

- Reducer unions the output of the mappers and presents rules:
  - <Key, Value> = <Walt:Jeff, reciprocal>
  - <Key, Value> <Walt:Stanley, directed>

The third reducer can shape the data any way that serves the business objective.



DELL EMC

The third time is the charm.

Mapper3 takes the combination of inbound [mail sent to] and outbound [mail sent from], and it outputs a key value pair of the form <Sender:Recipient>, relationship.

In this instance, the relationships are defined as either reciprocal, in which the sender sent/received mail to/from the recipient, or directed, in which a person sent mail to the recipient.

## Hadoop operational modes

### Hadoop operational modes

- Java MapReduce Mode
  - Write mapper, combiner, reducer functions in Java using Hadoop Java APIs
  - Read records one at a time
- Streaming mode
  - Uses \*nix pipes and standard input and output streams
  - Any language—Python, Ruby, C, Perl, Tcl/Tk, and so on
  - Input can be a line at a time, or a stream at a time

© 2018 Dell Inc. All rights reserved.

DELL.COM

Hadoop provides two operational modes, where each mode supports a type of interaction with MapReduce and HDFS. Since Hadoop is written in Java and provides Java classes and APIs to access them, **Java MapReduce mode** writes the mapper, combiner, and reducer functions in Java. In the Java MapReduce mode, input data is made available to each function, a record at a time.

In contrast, **streaming mode** supports standard \*nix streams—stdin, stdout—and the \*nix pipe mechanisms. This means that all the MapReduce functions can be written in any programming or scripting language you want—C, Ruby, Python, Perl, and so on. Although input can be read a line at a time, some languages support the “slurping” of the entire input stream into memory; Perl is one example.

Hadoop streaming reference:

[hadoop.apache.org/docs/current/hadoop-streaming/HadoopStreaming.html](http://hadoop.apache.org/docs/current/hadoop-streaming/HadoopStreaming.html)

## YARN—Yet Another Resource Negotiator

### YARN—Yet Another Resource Negotiator

- Yet Another Resource Negotiator (YARN) is a Hadoop ecosystem component that provides the resource management.
- YARN is also one of the most important components of the Hadoop ecosystem. YARN is called as the operating system of Hadoop, as it is responsible for managing and monitoring workloads.
- It allows multiple data processing engines such as real-time streaming and batch processing to handle data stored on a single platform.
- Main features of YARN are:
  - Flexibility
  - Efficiency
  - Shared
  - Scalability

BY DELL INC.

DELL INC.

Yet Another Resource Negotiator (YARN) is a Hadoop ecosystem component that provides the resource management. YARN is also one of the most important components of the Hadoop ecosystem. YARN is called the operating system of Hadoop, as it is responsible for managing and monitoring workloads. It allows multiple data processing engines, such as real-time streaming and batch processing, to handle data stored on a single platform.

The main features of YARN are:

- **Flexibility**

YARN enables other purpose-built data processing models beyond MapReduce—batch—such as interactive and streaming. Due to this feature of YARN, other applications can also be run along with MapReduce programs in Hadoop.

- **Efficiency**

As many applications run on the same cluster, efficiency of Hadoop increases without much effect on quality of service.

## Lesson: Introduction to advanced analytics—technology and tools

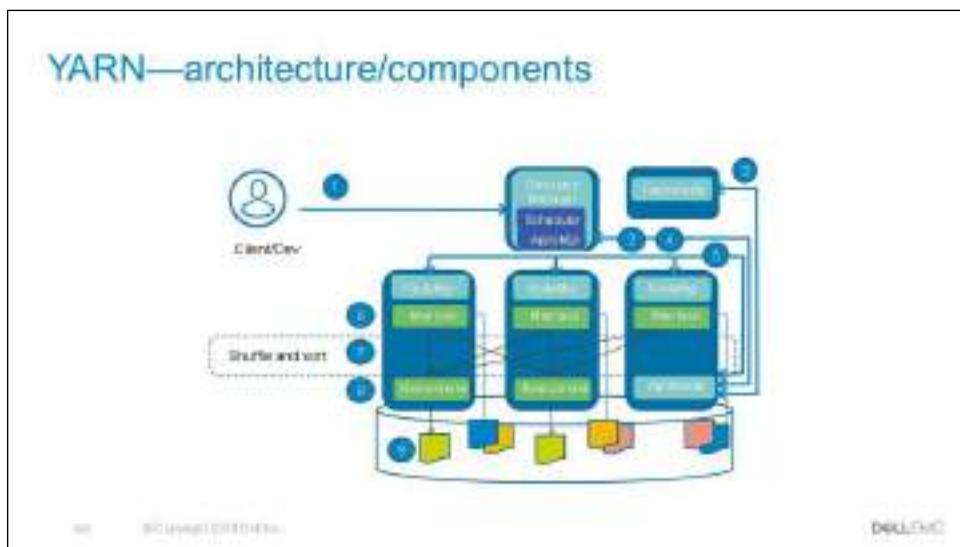
- **Shared**

Provides a stable, reliable, secure foundation and shared operational services across multiple workloads. Additional programming models such as graph processing and iterative modeling are now possible for data processing.

- **Scalability**

YARN supports very large clusters.

## YARN—architecture/components



Under YARN, the content and structure of a MapReduce job is unchanged, but the scheduling and management of the job is quite different. The JobTracker functionality is now shared by the ResourceManager and the ApplicationMaster, or AppMaster.

The key steps are:

1. The client submits a MapReduce job to the ResourceManager, which schedules the job based on cluster activity.
2. When the Scheduler decides to begin the MapReduce job, the ApplicationsManager—AppsMgr—starts the ApplicationMaster.
3. From the NameNode, the ApplicationMaster determines the nodes on which HDFS blocks are stored and builds an execution plan and resource requirements.
4. From the ResourceManager, the ApplicationMaster requests resources including RAM and specific node or rack names to minimize the transfer of data. The ResourceManager then informs the ApplicationMaster of the granted resources.
5. The ApplicationMaster instructs the NodeManagers—NodeMgr—to dedicate the allocated resources for each Map or Reduce task.
6. The ApplicationMaster starts the Map tasks and monitors their status.

## Lesson: Introduction to advanced analytics—technology and tools

7. The shuffle and sort occurs.
8. The ApplicationMaster starts the reduce tasks and monitors their status.
9. Finally, the reduce task output is written to HDFS. Also, the job-specific ApplicationMaster is shut down and its resources released.

## Check your knowledge

### Check your knowledge

1. Why is a combiner function optional in the MapReduce framework?
2. What is the purpose of the NameNode in HDFS?
3. Identify an embarrassingly parallel situation from your current work.
4. What are two benefits of YARN?



000 00000000000000000000000000000000

DELL EMC

## Check your knowledge



### Discussion

## Question / Discussion Topic:

1. Why is a combiner function optional in the MapReduce framework?
2. What is the purpose of the NameNode in HDFS?
3. Identify an embarrassingly parallel situation from your current work.
4. What are two benefits of YARN?

## Discussion Notes:

## Lesson—summary

### Lesson—summary

During this lesson, the following topics were covered:

- Applying the MapReduce/Hadoop Processing Framework in Big Data analytics problems for unstructured data
- Differentiating among the various elements of Hadoop
- Naming the components of HDFS
- Identifying the parts of a Hadoop streaming script



This lesson:

- Introduced the idea behind MapReduce processing, and then described how Hadoop implements this algorithm.
- Covered data management—the processing and development of frameworks to work on unstructured data in the terabyte range, and presented extensions to Hadoop that apply its capabilities.

## Lesson: Hadoop ecosystem

### Introduction

Lesson: Hadoop  
ecosystem



## Lesson: Hadoop ecosystem

### Lesson: Hadoop ecosystem

During this lesson, the following topics are covered:

- Use query languages—Hive and Pig—for data analytics problems using unstructured data.
- Build and query an HBase database.
- Suggest examples where HBase is most suitable.

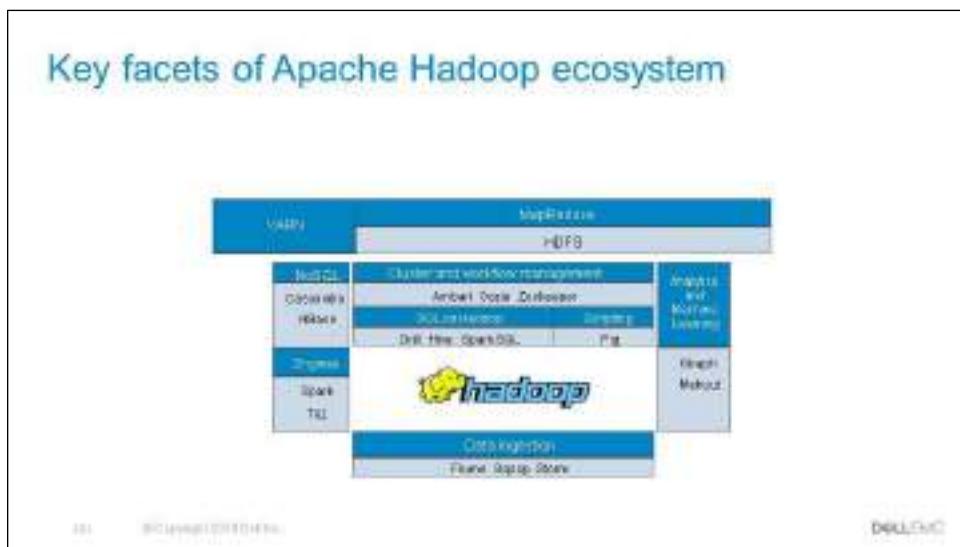


DELL EMC

Hadoop by itself was developed as an open-source implementation of Google, Inc.'s, Google File System (GFS) MapReduce framework. Further, Hadoop has spawned various associated projects, most of which depend on Hadoop's MapReduce capabilities and HDFS. (Exceptions are HBase and Cassandra.)

This lesson covers several of these.

## Key facets of Apache Hadoop ecosystem



This diagram provides several key categories for the Hadoop ecosystem members. At the center is Hadoop, which includes HDFS, MapReduce, and YARN. For the developers or analysts with experience in SQL, tools such as Drill, Hive, and Spark may be used to analyze HDFS data files.

Pig is a good choice for those users who are comfortable with scripting languages. For more complex data mining and analytical activities, tools such as Giraph and Mahout provide routines for graph and clustering analysis, for example.

More information about these tools and other Apache projects can be found at the following websites:

- [apache.org](http://apache.org)
- [incubator.apache.org](http://incubator.apache.org)

## Apache projects covered in this lesson

### Apache projects covered in this lesson

- **Pig**—Data flow language and execution environment
- **Hive**—and HiveQL—Query language based on SQL
- **HBase**—Column-oriented database
- **Spark**—Performs machine learning tasks in memory for faster and efficient performance
- **Mahout**—Highly scalable machine learning library designed to work on distributed systems

101

DATA SCIENCE

DELL EMC

These interfaces build on core Hadoop to support different styles of interaction with the data stored in HDFS. All are layered over HDFS and MapReduce.

**Pig** is a high-level programming language, useful for analyzing large datasets.

**Hive**—HiveQL—is an SQL-like query language for building MapReduce jobs.

**HBase** is a column-oriented database running over HDFS and supporting MapReduce and point queries.

HBase depends on **Zookeeper**, a coordination service for building distributed applications.

**Mahout** is a project to build scalable, machine learning systems by packaging machine learning algorithms as an executable library. You can specify input and output sources in HDFS to the algorithms—for example, each algorithm can take different parameters.

More information is available at [mahout.apache.org](http://mahout.apache.org).

## What is Pig?

### What is Pig?

- High-level data flow language for exploring large datasets
  - Can execute data flows in parallel in Hadoop
- Pig scripts are converted to MapReduce jobs by compiler
- Significantly reduces the code needed for execution
- Operates on files in HDFS
- Supports batch processing only
- Key properties of Pig
  - Ease of programming
  - Optimization opportunities
  - Extensibility

Dell EMC



Pig is a data-flow language and an execution environment to access the MapReduce functionality of Hadoop and HDFS.

It is a scripting language where the Pig compiler converts the code to MapReduce jobs and executes the tasks.

Parallelization is a key attribute for Pig and operates on HDFS systems. It will use existing metadata, if available, or can be run if it is not available, as well.

Properties of Pig are:

- **Ease of programming**—Can achieve parallel execution of simple and parallel data analysis tasks
- **Optimization opportunities**—Allows the user to focus on semantics rather than efficiency
- **Extensibility**—Allows users to create their own functions to do special-purpose processing
- **Code reduction**—On an average, reduces the code written to 5% of the MapReduce code\*

**A word of caution is in order:** If you only want to touch a small portion of a given dataset, then Pig is not for you, because it only knows how to read all the data

## Lesson: Hadoop ecosystem

presented to it. Pig only supports batch processing of data. So, if you need an interactive environment, Pig is not for you.

Pig can run in two execution environment modes:

- Local --- access a local file system
- MapReduce --- when you're interested in the Hadoop environment

\*Source: [hadooptutor.blogspot.com/2013/09/difference-between-mapreduce-and-pig.html](http://hadooptutor.blogspot.com/2013/09/difference-between-mapreduce-and-pig.html)

## Writing Pig Latin

### Writing Pig Latin

- A Pig script is a series of operations—transformations—applied to an input to produce an output.
  - Sample Pig command to load a tab delimited file
    - A = LOAD 'telephonesInformation.txt' using PigStorage ('\t') as (FName: chararray, LName: chararray, MobileNo: chararray, City: chararray, Profession: chararray)
    - B = FOREACH A generate FName, MobileNo, Profession;
    - DUMP B;
- Supports examining data structures and subsets of data
- Can execute Pig programs as a script
  - Via Grunt: an interactive shell or from a Java program
  - Via the command line: pig <scriptname>

Source: <https://databricks.com/>

DELL INC.

One can think of a data flow programming model as a series of transforms or filters applied to an input stream.

**In Pig, each transform is defined as a new input data source.** The next example presents a closer look at these operations.

Descriptions of these transforms can be provided to Pig via a Pig Latin script, or interactively using Grunt, Pig's command-line interface.

Grunt also provides commands to query intermediate steps in the process. EXPLAIN shows the related MapReduce plan, DUMP lists out a dataset, and DESCRIBE describes the schema structure for that particular dataset.

Someone described Pig in this way: “You can process terabytes of data by issuing a half-dozen lines of Pig Latin from the console.” Not a bad return on investment. Just ensure they are the right half-dozen lines.

## Apache Hive

### Apache Hive

- Hive is a data warehouse infrastructure built on top of Hadoop that can compile SQL queries and run the jobs in cluster.
- Data can be imported from RDBMS or from any other structured data sources (using tools like Apache Sqoop™).
- Hive provides the following features:
  - Tools to enable easy access to data via SQL.
  - A mechanism to impose structure on various data formats.
  - Hive is not an Online Transaction Processing (OLTP) tool; it is closer to Online Analytical Processing (OLAP).
  - Best suited for the applications where data is structured, static, and formatted.
  - Query execution via Apache Data Warehousing Tez™, Apache Spark™, or MapReduce.

DATA SCIENCE AND ANALYTICS

DELL EMC

The Hive system is aimed at the data scientist with strong SQL skills. Think of Hive as occupying a space between Pig and a DBMS, although that DBMS does not have to be a Relational DBMS {RDBMS}.

In Hive, all data is stored in tables. Hive manages the schema for each table. Tables can be populated via the Hive interface, or a Hive schema can be applied to existing data stored in HDFS.

Built on top of **Hadoop**, Hive provides the following features:

- Tools to enable easy access to data via SQL, thus enabling data warehousing tasks such as extract/transform/load (ETL), reporting, and data analysis
- A mechanism to impose structure on a variety of data formats
- Access to files stored either directly in **HDFS** or in other data storage systems such as **HBase**
- Query execution via Apache Tez™, Apache Spark™, or MapReduce

Online transaction processing (OLTP) is not yet supported in Hive, but Online Analytical Processing (OLAP) is supported, whereas in the traditional database, both OLTP and OLAP are supported.

Examples of when to use Hive:

- Data is imported into HDFS from a relational database

- Data is already well structured in a tabular format
- The analysts are familiar with SQL

## Hive Shell and HiveQL

### Hive Shell and HiveQL

- Hive
  - Provides web, server, and shell interfaces for clients
  - Hive shell is the default
    - Can run external host commands using “!prog” command
    - Can access HDFS using the DFS command
- HiveQL
  - Partial implementation of SQL-92—closer to MySQL.
  - Data in Hive can be in internal tables or external tables.
    - Hive manages the internal tables.
    - Hive does not manage external tables—lazy create and load.

111 / 800 pages of 1000 pages

Dell EMC

The Hive program provides different functions, depending on which commands are provided. The simplest invocation is simply “hive”, which brings up the Hive shell. From there, you can enter Hive SQL commands interactively, or these commands can be combined into a single script file.

*hive hwi* is the Hive web interface through which one can browse existing database schemas and create sessions for issuing database queries and Hive commands. The interface is available at <hivehost>:9999/hwi.

*Hive hiveserver* will start Hive as a server listening on port 10,000, which provides a Thrift and a JDBC/ODBC interface to Hive databases.

**Data for Hive can be stored in Hive’s internal tables—managed tables—or can be retrieved from data in the file system, HDFS.** An example of creation of an external table is:

```
CREATE EXTERNAL TABLE my_ext_data (dummy STRING) LOCATION  
'/opt/externalTable'  
LOAD DATA INPATH '/opt/externalTable' INTO TABLE my_ext_data
```

The existence of this data is not checked when these statements are executed, nor is data loading in Hive’s datastore. Hence, the notion of “lazy create and lazy load” which are examples of schema on read.

## Temperature example—Hive

The screenshot shows a presentation slide with a blue header bar containing the title 'Temperature example—Hive'. Below the title is a table with three rows, each containing a line of Hive code. The first row has a blue background and contains the text 'Example Hive code'. The second row contains line 1 of the code, which is 'CREATE TABLE records (year STRING, temperature INT, quality INT) ROW FORMAT DELIMITED FIELDS TERMINATED by \t;'. The third row contains line 2, 'LOAD DATA LOCAL 'data/samples/bf' OVERWRITE INTO TABLE records;'. Line 3 contains the SQL query 'SELECT year, MAX(temperature) FROM records WHERE temperature > 0 AND (quality == 0 OR quality == 1 OR quality == 4 OR quality == 5 OR quality == 8) GROUP BY year;'. At the bottom left of the slide, there is a small Dell logo and the text 'Data Science and Big Data Analytics v2'.

| Example Hive code  |
|--|
| 1 CREATE TABLE records (year STRING, temperature INT, quality INT) ROW FORMAT DELIMITED FIELDS TERMINATED by \t;   |
| 2 LOAD DATA LOCAL 'data/samples/bf' OVERWRITE INTO TABLE records;  |
| 3 SELECT year, MAX(temperature) FROM records WHERE temperature > 0 AND (quality == 0 OR quality == 1 OR quality == 4 OR quality == 5 OR quality == 8) GROUP BY year; |

Now, consider the example of calculating the maximum temperature for a given year from thousands and thousands of weather observations, from hundreds of weather stations.

Line 1 defines your table and states that your input consists of tab-delimited fields.

In line 2, you encounter LOAD DATA again, with a slightly different syntax.

Line 3 is a standard SQL query that produces a relation consisting of a year and the max temperature for that year. The ROW FORMAT clause is a Hive-specific addition.

Hive maintains its own set of tables; these tables could exist on a local file system or in HDFS as /usr/hive/XXXXX. A directory in the file system corresponds to a particular table.

Hive does not implement the full SQL-92 standard, and provides certain clauses that do not appear in standard SQL. "ROW FORMAT ..." is one such example.

## Hive comparison with SQL

| Hive   | Database                                  |
|--|---|
| "Schema on read"                                     | "Schema on write"                         |
| Easily scalable at low cost                          | Expensive scaling                         |
| Follows notation: write once, read many times.       | Can read or write as many times as needed |
| Incomplete SQL-92—never a design goal                | Full SQL-92                               |
| No updates, transactions, indexes available in v0.7+ | Updates, transactions, and indexes        |

### Comparing Hive and SQL

Hive → Schema on READ: it does not verify the schema while it is loading the data.

Traditional database → Schema on **write**: table schema is enforced at data load time—that is, if the data being loaded does not conform on schema, in which case it is rejected.

Hive is very easily scalable at low cost. Traditional database is not very scalable and has a costly scale-up.

Hive is based on Hadoop notation that is write-once and read-many-times. In a traditional database, you can read and write many times.

Hive → Record-level updates are not possible in Hive. Traditional database → Record-level updates, insertions and deletes, transactions, and indexes are possible.

In most traditional DBMS, the database description—or schema—is read and applied when the data is loaded. If the data does not match the schema—specifically, the table into which it is read—then the load fails. This is often called “Schema on write.”

Hive, on the other hand, does not attempt to apply the schema until the data is actually read when someone issues a query. This results in fast loads, and supports multiple schemas for the same data—only defining as many variables as needed for your particular analysis. In addition, the actual format of the data may not be known because queries against the data have not been defined.

Updates and transactions aren't supplied with Hive. Indexes are available as of Hive 0.8. If you want concurrent access to tables, then the application must roll its own. That being said, the Hive project is working toward integration with the HBase project that does provide row updates.

See the Hive project page at [hive.apache.org](http://hive.apache.org) for further details.

## HBase—the Hadoop database from Apache

### HBase—the Hadoop database from Apache

- Apache HBase is a popular and highly efficient Column-oriented NoSQL database built on top of the Hadoop Distributed File System. It allows for performing read/write operations on large datasets in real time using Key/Value data.\*
- HDFS is good for batch processing.
  - Not good for record lookup
  - Not good for updates
- It is an open-source, distributed, versioned, column-oriented store modeled after Google's Bigtable.
- HBase addresses these problems.
- It uses Zookeeper extensively for region assignment.

10

DATA SCIENCE

DELL EMC

Source: [intellipaat.com/blog/what-is-apache-hbase](http://intellipaat.com/blog/what-is-apache-hbase)

Apache HBase is a NoSQL database that runs on top of Hadoop as a distributed and scalable Big Data store. This means that HBase can apply the distributed processing paradigm of the Hadoop Distributed File System and benefit from Hadoop's MapReduce programming model.

HBase represents a further layer of abstraction on Hadoop. HBase has been described as “a distributed column-oriented database [data storage system]” built of top of HDFS.

HBase is described as managing *structured* data. Each record in the table can be described as a key—treated as a byte stream—and a set of variables, each of which may be versioned. It is not structured in the same sense as an RDBMS is structured.

HBase is a more complex system than those seen previously. HBase uses more Apache Foundation open-source frameworks. Zookeeper is used as a coordination system to maintain consistency, Hadoop for MapReduce and HDFS, and Oozie for workflow management.

Zookeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group service.

Reference: [Zookeeper.apache.org](https://zookeeper.apache.org)

## When to choose HBase

### When to choose HBase

- You need random, real-time read/write access to your Big Data.
- You need Schema-on-write:
  - When new attributes may be added as new data is processed (new words).
- You need sparse tables consisting of billions of rows and millions of columns where each column variable may be versioned.
- Google's Bigtable: a "web table."
- HBase as a schema-less data store; that is, it is fluid—you can add to, subtract from, or modify the schema as you go along.
  - If you wanted to perform text analysis on the Google n-gram corpus, you have one trillion words with 1 million distinct words. The frequency table such as this is ideally suited for HBase.

100

DATA SCIENCE

DELL EMC

HBase has been described in this way: “[its] forte is when real-time read-write random access to very large datasets is required.” HBase is an open-source version of Google’s Bigtable, and it is instructive to read the definition of BigTable by the original authors:

*BigTable is a distributed storage system for managing structured data that is designed to scale to a very large size: petabytes of data across thousands of commodity servers.*

Because HBase is not constrained in the same way as an RDBMS system is constrained, HBase designs can take advantage of the physical layout of the table on disk, to increase performance.

It is useful to recall that Google’s Bigtable was designed to store information about web URLs—web documents. Fields in this table could be versioned—new versions of an .HTML page, for example—and the table could be updated frequently as web-crawlers discovered new data. One design decision to speed access to URLs from the same site was to reverse the order of the URL: instead of media.Google.com, the URL would be stored as com.Google.media, ensuring that other Google URLs would be reasonably close.

## HBase comparison with traditional database

| Hive External Tables                         | Hive Internal Tables                      | Databases                                 |
|--|---|---|
| “Schema on read”                             | “Schema on write”                         | “Schema on write”                         |
| Easily scalable at low cost                  | Easily scalable at low cost               | Expensive scaling                         |
| Follows insertion order once read many times | Can read or write as many times as needed | Can read or write as many times as needed |
| Incomplete SQL-92                            | Closer to SQL-92                          | Full SQL-92                               |
| No updates, transactions, or indexes only    | No transactions, updates and indexes only | Updatable, transactions, and indexes      |

000      © 2018 Dell Inc.      DELL.COM

Although HBase may appear to be a traditional DBMS, it is not.

HBase is a distributed, column-oriented data storage system that can scale billions-of-rows tall, millions-of-columns wide, and can be horizontally partitioned and replicated across thousands of commodity servers automatically.

The HBase table schemas mirror physical storage for efficiency; an RDBMS does not. The RDBMS schema implies no specific physical structuring.

HBase is purpose built for one table, and it can read and write data based on a specific key. RDBMS, can have any number of tables with many primary and secondary keys.

Most RDBMS systems require that data must be consistent after each transaction—ACID properties. Not Only SQL (NoSQL) systems such as HBase do not suffer from these constraints, and implement strong consistency. This means that, for some systems, you cannot write a value into the database and immediately read it back in. Strange, but true.

Another of HBase’s strengths is in its wide-open view of data is that HBASE will accept almost anything it can cram into an HBase table.

## Mahout

The screenshot shows the official Mahout project page. At the top, the word "Mahout" is written in a large, blue, sans-serif font. Below it, a sub-header reads "Scalable machine learning and data mining library for Hadoop". Underneath this, another sub-header says "Example applications use cases". Four square icons are displayed, each with a small image and a label: "Recommendation mining" (showing a user profile and items), "Classification" (showing a shopping cart), "Clustering" (showing three colored dots), and "Regression" (showing a bar chart). At the bottom of the page, there are links for "HOME", "ABOUT", "CONTACT", and "DELL.COM".

Mahout is a set of machine learning algorithms that applies Hadoop to provide both data storage.

The mahout command is itself a script that wraps the Hadoop command and executes a requested algorithm from the Mahout job jar file—jar files are Java ARchives, and are very similar to Linux tar files, tape archives. Parameters are passed from the command line to the class instance.

If you plan on using Mahout, remember that these distributions—Hadoop and Mahout—anticipate running on a nix machine, although a Cygwin environment on Windows will work, too—or rewriting the command scripts in another language, say as a batch file on Windows. It goes without saying that a compatible working version of Hadoop is necessary. Lastly, Mahout requires that you program in Java—no other interface outside of the command line is supported.

## Examples of useful algorithms available in Mahout

Mahout provides various different algorithms to support its use cases.

Check the website at [mahout.apache.org](http://mahout.apache.org) for more details.

## Apache Spark™

### Apache Spark™

- Analytics engine for large-scale data processing
- High performance for both batch and streaming data
  - Near real-time streaming data support
- Combines SQL, streaming, and complex analytics
  - SQL and DataFrames
  - MLlib (for machine learning)
  - GraphX
  - Spark Streaming
- Runs on many platforms
  - Standalone cluster mode
  - Hadoop YARN
  - Mesos and Kubernetes

Apache Spark™ is a unified analytics engine for large-scale data processing. Unlike HDFS, which is suitable for batch processing, Spark achieves high performance for both batch and streaming data.

Spark has many features:

- Speed

Near real-time streaming data support

In-memory

- Ease of use

You can write applications for Spark in Java, Scala, Python, R, and SQL. You may also use it interactively in the scripting language shells.

- Generality

Spark Combines SQL, streaming, and complex analytics for you to create applications using all these features.

They include:

- SQL and DataFrames
- MLlib (for machine learning)

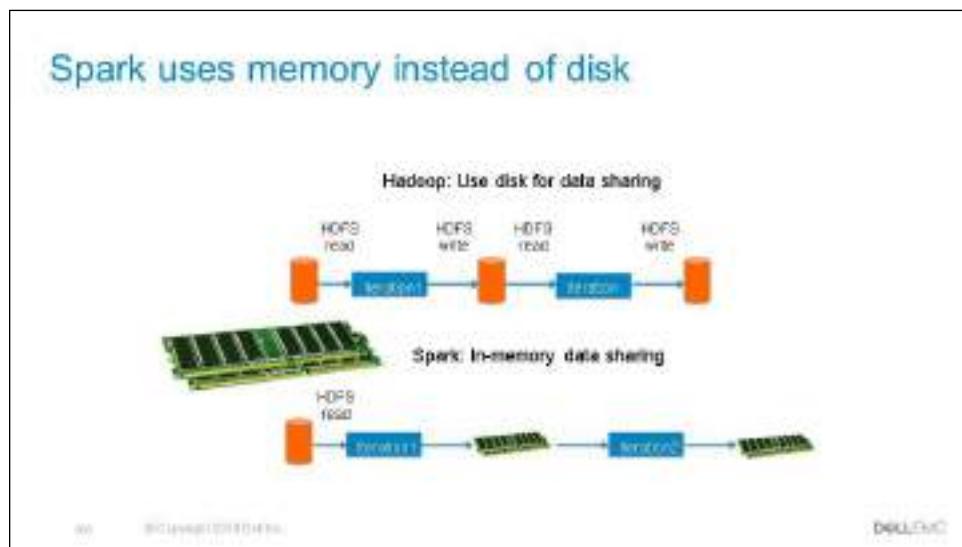
- GraphX (for graphs)
- Spark Streaming
- Wide platform support

You can run Spark using its standalone cluster mode, on EC2, on Hadoop YARN, on Mesos, or on Kubernetes.

You may access data in HDFS, Apache Cassandra, Apache HBase, Apache Hive, and dozens of other data sources.

For the latest features please visit: [spark.apache.org](http://spark.apache.org).

## Spark uses memory instead of disk



**HDFS**—Every time an operation is performed in HDFS using MapReduce, first the data is read from the cluster, then the operation or iteration is performed, and the data is rewritten into the cluster. It is time consuming to complete this operation. So, Dell started using Spark.

**Spark**—Using this, Dell significantly reduces the processing time to perform an operation in Hadoop. Spark helps run operations by saving data in a memory cache and performing iterations, and it runs all the steps faster and more efficiently.

## Sort competition

|                          | Hadoop MR<br>Running on HDFS           | Apache Spark<br>Running on HDFS   |
|--------------------------|--|-----------------------------------|
| Duration                 | 1823 TB                                | 108 TB                            |
| Elapsed time             | 72 mins                                | 23 mins                           |
| Nodes                    | 2108                                   | 286                               |
| RAM                      | 51408 GB total                         | 6500 MB total                     |
| Churn rate<br>throughput | 315.09GB/sec                           | 819.09GB/sec                      |
| Hardware                 | Decades old—DC2—<br>10.00GBs available | Modern—DC2—<br>10.00GBs available |
| <b>Sort time</b>         | <b>1.42 TB/min</b>                     | <b>4.27 TB/min</b>                |
| <b>Sort吞吐量</b>           | <b>9.67 GB/min</b>                     | <b>26.7 GB/min</b>                |

Spark, 2x faster with 1/10 the nodes

Sort benchmark: Daybirra Dayal's sort of 100 TB of ints—1 trillion records  
[databricks.com/blog/2014/11/06/spark-chockily-sorts-a-new-record-in-the-gcsort-sorting-test/](http://databricks.com/blog/2014/11/06/spark-chockily-sorts-a-new-record-in-the-gcsort-sorting-test/)

Here is an example of comparing a task run using Hadoop MapReduce and Spark. As seen here, Spark is more than three times faster, compared to MapReduce.

## Check your knowledge

### Check your knowledge

1. How does Pig differ from a typical MapReduce process?
2. How does schema parsing differ in Hive external tables from a traditional RDBMS?
3. With regards to file structure, how does HBase differ from a traditional RDBMS?



000 - 00000000000000000000000000000000

DELL EMC

## Check your knowledge



### Discussion

## Question / Discussion Topic:

1. How does Pig differ from a typical MapReduce process?
2. How does schema parsing differ in Hive external tables from a traditional RDBMS?
3. With regards to file structure, how does HBase differ from a traditional RDBMS?

## Discussion Notes:

## Lesson—summary

### Lesson—summary

During this lesson, the following topics were covered:

- Query languages for Hadoop—Hive and Pig
- HBase—a BigTable workalike using Hadoop
- Mahout—machine learning algorithms using Hadoop MapReduce and HDFS
- Other elements of the Hadoop ecosystem



000

Dell EMC

Dell EMC

This lesson covered:

- Hive and Pig—Hadoop query languages
- HBase—a BigTable workalike using Hadoop
- Mahout—machine learning algorithms and Hadoop MapReduce

## Lesson: In-database analytics SQL essentials

### Introduction

Lesson: In-database  
analytics SQL essentials



## Lesson 3—in-database Analytics SQL essentials

### Lesson 3—in-database Analytics SQL essentials

During this lesson, the following topics are covered:

- SET Operations
- Online analytical processing (OLAP) features
- GROUPING SETS, ROLLUP, CUBE
- GROUPING, GROUP\_ID functions
- Text processing, Pattern matching

These topics are covered in this lesson.

## Analytical databases and libraries

### Analytical databases and libraries

- This lesson introduces PostgreSQL, Greenplum, and MADlib.
- PostgreSQL—It is an object-relational database management system. Features such as Constraints, Triggers, and rules make it object-relational. PostgreSQL is an RDBMS with object-oriented features.
- Greenplum Database—It is an advanced, fully featured, open-source data platform. It provides powerful and rapid analytics on petabyte scale data volumes.
- MADlib—Apache™ MADlib® is an open-source library for scalable in-database analytics. It provides data-parallel implementations of machine learning, mathematical, and statistical methods on the Greenplum database, PostgreSQL, and others.

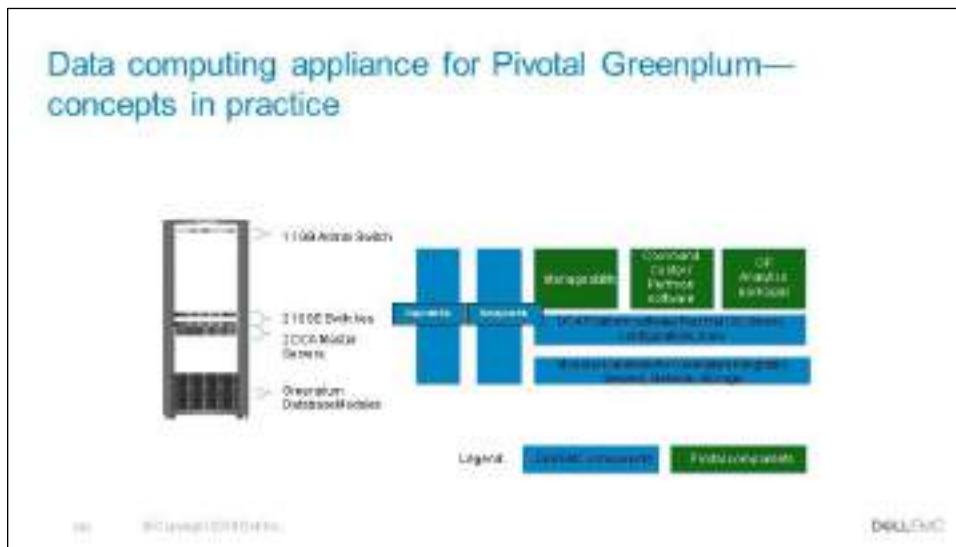


GREENPLUM  
DATABASE



DELL EMC

## Data computing appliance for Pivotal Greenplum—concepts in practice



To enable the successful deployment of a Greenplum database, Dell EMC offers the Data Computing Appliance (DCA). The DCA consists of integrated servers, networking, and storage. As the database grows, additional Greenplum database modules can be added. The DCA platform software can manage hardware configurations from one-fourth of a rack to 11 racks. For more details, go to [pivotal.io/emc-dca](http://pivotal.io/emc-dca).

## Set operations

### Set operations

As an ANSI standard SQL database, Greenplum supports the following set operations as part of a SELECT statement:

- INTERSECT—Returns rows that appear in all answer sets.
- EXCEPT—Returns rows from the first answer set and excludes those rows from the second.
- UNION ALL—Returns a combination of rows from multiple SELECT statements with repeating rows.
- UNION—Returns a combination of rows from multiple SELECT statements with no repeating rows.

14

© Copyright 2018 Dell Inc.

DELL.COM

## Set operations

Set operators:

- Manipulate the results sets of two or more queries by combining the results of individual queries into a single result set.
- Do not perform row-level filtering.

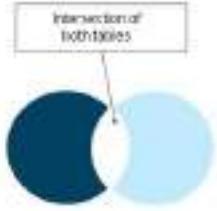
Set operations supported by Greenplum (an ANSI standard SQL database) are:

- INTERSECT, which returns rows that appear in all answer sets generated by individual SELECT statements.
- EXCEPT, which returns all rows from the first SELECT except for those that also selected by the second SELECT. This operation is the same as the MINUS operation.
- UNION ALL, which combines all the results of two or more SELECT statements. There may be repeating rows.
- UNION, which combines the results of two or more SELECT statements. There will be no repeating rows.

## Set operations—INTERSECT

### Set operations—INTERSECT

**INTERSECT:**  
Returns only the rows that appear in both SQL queries.  
Removes duplicate rows.



The diagram shows two overlapping circles. The overlapping area is shaded blue and labeled "Intersection of both tables". A callout box points to this intersection area with the text "Intersection of both tables".

Find all the suppliers in supplier tables that have had orders where the quantity more than 0.

```
SELECT supplier_id  
FROM suppliers  
INTERSECT  
SELECT supplier_id  
FROM orders  
WHERE quantity>0;
```

DELL EMC

## Set operations—INTERSECT

A set operation takes the results of two queries and returns only the results that appear in both result sets. Duplicate rows are removed from the final set returned.

## Set Operations—EXCEPT

### Set Operations—EXCEPT

**EXCEPT:**  
Returns all rows from the first SELECT statement  
Omits all rows that appear in the second SELECT statement

Diagram illustrating the EXCEPT operation: Two tables, Table A and Table B, are shown. A self-join arrow connects Table A to itself, pointing from the 'Excluded' row back to the original 'Excluded' row. This visualizes how the EXCEPT operation retains rows unique to the first query and excludes rows found in the second query.

```
SELECT * FROM TableA WHERE total_gear > 100
EXCEPT
SELECT * FROM TableB WHERE total_gear > 100
```

DELL EMC

### Set operations—EXCEPT

The EXCEPT set operation takes the distinct rows of the first query and returns all of the rows that do not appear in the result set of the second query.

## Set Operations—UNION ALL

**Set Operations—UNION ALL**

**UNION ALL:**  
Combines rows from the first query with rows from the second query  
Does not remove duplicate rows

All rows from both tables

|                          |
|--------------------------|
| 10001 Company A Inc.     |
| 10002 Apple Inc.         |
| 10003 Amazon.com Inc.    |
| 10004 Facebook Inc.      |
| 10005 Google Inc.        |
| 10006 Microsoft Corp.    |
| 10007 Oracle Corp.       |
| 10008 IBM Corp.          |
| 10009 Dell Inc.          |
| 10010 Cisco Systems Inc. |

Dell Inc.

### Set operations—UNION ALL

The UNION ALL set operation is similar to the UNION operation, but it does not remove duplicate or repeating rows.

## Set Operations—UNION

### Set Operations—UNION

**UNION:**  
Combines rows from the first query with rows from the second query  
Removes duplicates or repeating rows

```
SELECT * FROM TableA WHERE id = 1000
SELECT * FROM TableB WHERE id = 1000
SELECT * FROM TableA WHERE id = 1000
SELECT * FROM TableB WHERE id = 1000
```

Dell EMC

### Set operations—UNION

A UNION operation combines the results of the SELECT statement from the first table with the results from the query on the second table. The result set does not contain any repeating rows.

## Greenplum SQL OLAP grouping extensions

### Greenplum SQL OLAP grouping extensions

Greenplum supports the following grouping extensions:

- Standard GROUP BY
- ROLLUP
- GROUPING SETS
- CUBE
- grouping(column [, ...]) function
- group\_id() function

140

BIG DATA AND ANALYTICS

DELL EMC

## Greenplum SQL OLAP grouping extensions

Greenplum introduced support for extensions to the standard GROUP BY clause, which is fully supported. These clauses can simplify the expression of complex groupings:

- ROLLUP—This extension provides hierarchical grouping.
- CUBE—Complete cross-tabular grouping, or all possible grouping combinations, is provided with this extension.
- GROUPING SETS—Generalized grouping is provided with the GROUPING SETS clause.
- grouping function—This clause helps identify super-aggregated rows from regular grouped rows.
- group\_id function—This clause is used to identify duplicate rows in grouped output.

## Standard GROUP BY example

| product_name | quantity | row_id |
|--------------|----------|--------|
| apple        | 10       | 1      |
| apple        | 10       | 2      |
| apple        | 10       | 3      |
| apple        | 10       | 4      |
| apple        | 10       | 5      |
| apple        | 10       | 6      |
| apple        | 10       | 7      |
| apple        | 10       | 8      |
| apple        | 10       | 9      |
| apple        | 10       | 10     |
| apple        | 10       | 11     |
| apple        | 10       | 12     |
| apple        | 10       | 13     |
| apple        | 10       | 14     |
| apple        | 10       | 15     |
| apple        | 10       | 16     |
| apple        | 10       | 17     |
| apple        | 10       | 18     |
| apple        | 10       | 19     |
| apple        | 10       | 20     |
| apple        | 10       | 21     |
| apple        | 10       | 22     |
| apple        | 10       | 23     |
| apple        | 10       | 24     |
| apple        | 10       | 25     |
| apple        | 10       | 26     |
| apple        | 10       | 27     |
| apple        | 10       | 28     |
| apple        | 10       | 29     |
| apple        | 10       | 30     |
| apple        | 10       | 31     |
| apple        | 10       | 32     |
| apple        | 10       | 33     |
| apple        | 10       | 34     |
| apple        | 10       | 35     |
| apple        | 10       | 36     |
| apple        | 10       | 37     |
| apple        | 10       | 38     |
| apple        | 10       | 39     |
| apple        | 10       | 40     |
| apple        | 10       | 41     |
| apple        | 10       | 42     |
| apple        | 10       | 43     |
| apple        | 10       | 44     |
| apple        | 10       | 45     |
| apple        | 10       | 46     |
| apple        | 10       | 47     |
| apple        | 10       | 48     |
| apple        | 10       | 49     |
| apple        | 10       | 50     |
| apple        | 10       | 51     |
| apple        | 10       | 52     |
| apple        | 10       | 53     |
| apple        | 10       | 54     |
| apple        | 10       | 55     |
| apple        | 10       | 56     |
| apple        | 10       | 57     |
| apple        | 10       | 58     |
| apple        | 10       | 59     |
| apple        | 10       | 60     |
| apple        | 10       | 61     |
| apple        | 10       | 62     |
| apple        | 10       | 63     |
| apple        | 10       | 64     |
| apple        | 10       | 65     |
| apple        | 10       | 66     |
| apple        | 10       | 67     |
| apple        | 10       | 68     |
| apple        | 10       | 69     |
| apple        | 10       | 70     |
| apple        | 10       | 71     |
| apple        | 10       | 72     |
| apple        | 10       | 73     |
| apple        | 10       | 74     |
| apple        | 10       | 75     |
| apple        | 10       | 76     |
| apple        | 10       | 77     |
| apple        | 10       | 78     |
| apple        | 10       | 79     |
| apple        | 10       | 80     |
| apple        | 10       | 81     |
| apple        | 10       | 82     |
| apple        | 10       | 83     |
| apple        | 10       | 84     |
| apple        | 10       | 85     |
| apple        | 10       | 86     |
| apple        | 10       | 87     |
| apple        | 10       | 88     |
| apple        | 10       | 89     |
| apple        | 10       | 90     |
| apple        | 10       | 91     |
| apple        | 10       | 92     |
| apple        | 10       | 93     |
| apple        | 10       | 94     |
| apple        | 10       | 95     |
| apple        | 10       | 96     |
| apple        | 10       | 97     |
| apple        | 10       | 98     |
| apple        | 10       | 99     |
| apple        | 10       | 100    |

## Standard GROUP BY example

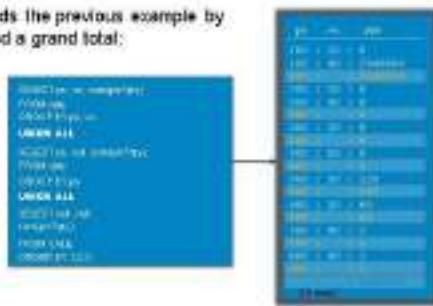
The standard GROUP BY clause groups results based on one or more columns specified. It is used in conjunction with aggregate statements, such as SUM, MIN, or MAX. This helps to make the resulting dataset more readable.

The slide shows an example of a standard GROUP BY clause used to summarize product sales by vendor.

## Standard GROUP BY example with UNION ALL

**Standard GROUP BY example with UNION ALL**

This example extends the previous example by adding subtotals and a grand total:



The image shows a Microsoft Word document window. On the left, there is a code block containing SQL-like pseudocode. On the right, there is a screenshot of a database management system interface showing a table with data grouped by category.

### Standard GROUP BY example with UNION ALL

In this follow-up example, the requirements for the query have been extended to include subtotals and a grand total. You would use a UNION ALL to continue the grouping and provide for the additional requirements.

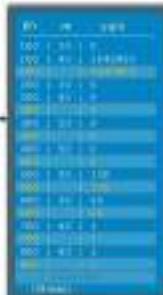
## ROLLUP example

### ROLLUP example

The following example meets the requirement of including the subtotal and grand totals:

```
SELECT * FROM Sales  
GROUP BY ROLLUP(Region, Month)  
ORDER BY ROLLUP.
```

As seen from the last slide, the result from both queries are the same. The same result can also be obtained using GROUPING SETS, as seen in the next slide.

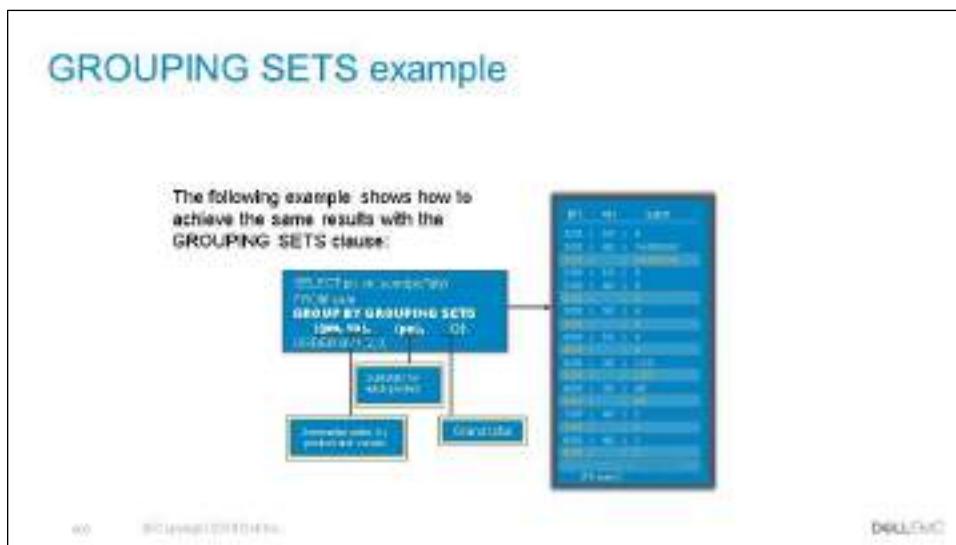


DATAVIZ

## ROLLUP example

This slide meets the requirements provided in the previous slide, but it uses the ROLLUP grouping extension. ROLLUP allows you to perform hierarchical grouping and helps to reduce the code.

## GROUPING SETS example



## GROUPING SETS example

The GROUPING SETS extension allows you to specify grouping sets. If you use the GROUPING SETS clause to meet the earlier requirements so that it produces the same output as ROLLUP, it would use the following groups:

- (pn,vn)—This grouping summarizes product sales by vendor.
- (pn)—This grouping provides subtotal sales for each product.
- ()—This grouping provides the grand total for all sales for all vendors and products.

## CUBE example

**CUBE example**

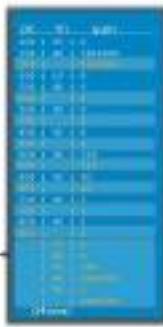
CUBE creates subtotals for all possible combinations of grouping columns.

The following example

```
SELECT *  
FROM Sales  
GROUP BY SalesDate  
ORDER BY SalesDate;
```

is the same as

```
SELECT *  
FROM Sales  
GROUP BY GROUPING SETS  
(SalesDate, SalesDate)  
ORDER BY SalesDate;
```



DATAVIEW

## CUBE example

A CUBE grouping creates subtotals for all of the possible combinations of the given list of grouping columns, or expressions.

In terms of multidimensional analysis, CUBE generates all the subtotals that could be calculated for a data cube with the specified dimensions.

In the example shown here, the additional grouping set of (vn), subtotaling the sales by vendor, is included as part of the cube.

**Note** that n elements of a CUBE translate to  $2^n$  (2 to the power of n) grouping sets. Consider using CUBE in any situation requiring cross-tabular reports. CUBE is typically most suitable in queries that use columns from multiple dimensions rather than columns representing different levels of a single dimension. For instance, a commonly requested cross-tabulation might need subtotals for all the combinations of month, state, and product.

## GROUPING function example

The screenshot shows a Microsoft SQL Server Management Studio window with two result sets. The top result set is a table with columns: Store, CustomerID, ProductID, and Total. One row for store 2 has a Total of 44. The bottom result set is the same table, but it includes an additional column labeled 'GROUPING\_ID'. For the row with Total 44, the GROUPING\_ID is 1, while for other rows it is 0. This visual comparison demonstrates how the GROUPING function can be used to distinguish between summary rows and regular data rows.

## GROUPING function example

When you use grouping extensions to calculate summary rows, such as subtotals and grand totals, the output can become confusing if the data in a grouping column contains NULL values. It is hard to tell if a row is supposed to be a subtotal row or a regular row containing a NULL.

In the example shown here, one of the rows shows a customer field that is NULL. Without the grouping ID, you could misinterpret the sum of 44 as a subtotal row for store 2.

The GROUPING function returns a result for each output row, where:

- 1 represents a summary row.
- 0 represents the others.

## GROUP\_ID function

The screenshot shows a section titled "GROUP\_ID function" with the following text:

**GROUP\_ID:**  
Returns 0 for each output row in a unique grouping set.  
Assigns a serial number >0 to each duplicate grouping set found.  
Can be used to filter output rows of duplicate grouping sets, such as in the following example:

```
SELECT a,b,c,group_id
  FROM t1
 GROUP BY ROLLUP(a,b,c)
 HAVING group_id > 0
 ORDER BY group_id;
```

Below the code, there are two small icons: a blue square with "400" and a green square with "DBMS\_STATS".

**GROUP\_ID function** Is useful when combining grouping extension clauses.

In this example query, the combination of ROLLUP and CUBE produces:

- 12 grouping sets.
- 8 DISTINCT grouping sets.

The group\_id function can be used to filter out or identify duplicate grouping sets in the output.

GROUP BY ROLLUP (a,b), CUBE (b,c) is the same as:

- GROUP BY GROUPING SETS ( (a,b), (a), () ), GROUPING SETS ( (b,c), (b), (c), () ).
- GROUP BY GROUPING SETS ((a,b,b,c), (a,b,b), (a,b,c), (a,b), (a,b,c), (a,b), (a,c), (a), (b,c), (b), (c), () )

Where there are 12 total grouping sets but only 8 distinct grouping sets, where the groups are:

- (a,b,b,c) = (a,b,c) = (a,b,c)
- (a,b,b) = (a,b) = (a,b)
- (a,c)
- (b,c)

## Lesson: In-database analytics SQL essentials

- (a)
- (b)
- (c)
- ()

## In-database text analysis

### In-database text analysis

- SQL features for
  - Text handling functions
  - Pattern matching with regular expressions
- Example use-cases
  - Filter emails with spam tag in subject
  - Extract domains from a URL
  - Extract all URLs from an .HTML file
  - Check for syntactically correct email addresses
  - Convert 10 digits into format "(123) 456-7890"

496

SQL Fundamentals

Dell EMC

## Pattern matching—regular expressions (Regex)

Pattern matching—regular expressions (Regex)

Regular expression match operators

| Operator        | Description                | Example                          |
|-----------------|----------------------------|----------------------------------|
| <code>~</code>  | Case-sensitive substring   | <code>'Greenplum' ~ 'ree'</code> |
| <code>~*</code> | Case-insensitive substring | <code>'Greenplum' ~* 'ee'</code> |

SQL functions

- `substring(string, from, pattern [for escape])`
- `regexp_matches(string, pattern, [Flags])`
- `regexp_replace(string, pattern, repl, [Flags])`
- `regexp_split_to_array(string)`

40 / 80 | © 2018 Dell Inc.

Dell EMC

The substring function is primarily used for string pattern matching.

The operator `~` is specified for case-sensitive match of the substring, and `~*` is specified for case-insensitive match.

In the first example, you are trying to find records with substring “Green” (case sensitive), starting at the beginning specified with character `^`

In the second example, you are finding a match for a pattern “ee” (case insensitive) as a preceding term one or more times, specified with a `+`

See [www.postgresql.org/docs/8.3/static/functions-matching.html](http://www.postgresql.org/docs/8.3/static/functions-matching.html) for details of the syntax.

## Regular expression quantifiers

| Regular expression quantifiers |  |
|--------------------------------|--|
| Quantifier                     | Matched                                  |
| .                              | Arbitrary character                      |
| ^ And \$                       | Virtual characters for beginning and end |
| *                              | Preceding item zero or more times        |
| +                              | Preceding item one or more times         |
| ?                              | Preceding item is optional               |
| {n}                            | Preceding item n times                   |
| {n,m}                          | Item, n to m                             |

00: REGULAR EXPRESSIONS  
DELL EMC

Quantifiers specify how often the preceding regular expression should match.

- Try to match the preceding regular expression zero or more times.
  - Example: "(ab)c\*" matches "ab" followed by zero or more "c"s, that is, "ab", "abc", "abcc", "abccc" ...
- Try to match the preceding regular expression one or more times.
  - Example: "(ab)c+" matches "ab" followed by one or more "c"s, that is, "abc", "abcc", "abccc" ...

Examples:

- All mail with at least two + in x\_spam\_level:

```
SELECT * FROM mail
WHERE x_spam_level ~ '\+\+\+*
```

- All top-level domains of sender's addresses:

```
SELECT substring("from" FROM
'\.[\[:alnum:]\]+\$') FROM mail
```

- Remove [Spam] tag at beginning of subjects:

```
SELECT regex_replace(subject,
'^((?:Re:[\[:space:]\])*|[Spam\\]|' || '(.*)', '\\1\\2') FROM mail
```

## Check your knowledge

### Check your knowledge

- How would you use GROUPING SETS to produce the same results as the following GROUP BY CUBE?

```
SELECT state, productID,  
SUM(volume) FROM sales GROUP  
BY CUBE (state, productID) ORDER  
BY state, productID.
```

- How would you show the subtotals for each week, for each state, and for each product? No other totals or grand totals are needed. Suppose that the table structure is TABLE sales (productID VARCHAR, state CHAR(2), week DATE, volume INT).



Note your answer/references below:

- SELECT state, productID, SUM(volume) FROM sales GROUP BY GROUPING SETS ((state, productID), (state), (productID), ()) ORDER BY state, productID.
- SELECT state, productID, week, SUM(volume) FROM sales GROUP BY GROUPING SETS ((state), (productID), (week)) ORDER BY state, productID, week.

## Lesson—summary

### Lesson—summary

During this lesson, the following SQL Essentials topics were covered:

- OLAP features
- GROUPING SETS, ROLLUP, CUBE
- GROUPING, GROUP\_ID functions
- Text processing, pattern matching



400

SQL Fundamentals

Dell EMC

SQL essentials were covered in this lesson.

## Lesson: Advanced SQL and MADlib

### Introduction

Lesson: Advanced SQL  
and MADlib



## Lesson—advanced SQL and MADlib

### Lesson—advanced SQL and MADlib

During this lesson, the following topics are covered:

- Window functions
- User-defined functions and aggregates
- Ordered Aggregates
- MADlib

This lesson covers advanced SQL and MADlib functions.

## Window functions

A window function performs a calculation across a set of rows that are related to the current row. This is similar to what can be completed with the aggregate function.

Unlike traditional aggregate functions, window functions do not cause grouping of rows to a single output.

Window functions can access other rows based on partition and order criteria that the user specifies.

A window function performs a calculation across a set of table rows that are somehow related to the current row. This is comparable to the type of calculation that can be completed with an aggregate function. But, unlike regular aggregate functions, use of a window function does not cause rows to become grouped into a single output row. The rows retain their separate identities. Behind the scenes, the window function is able to access more than just the current row of the query result.

Window functions allow application developers to more easily compose complex OLAP queries using standard SQL commands. For example:

- Moving averages or sums can be calculated over various intervals.
  - Aggregations and ranks can be reset as selected column values change.
  - Complex ratios can be expressed in simple terms.

Window functions can only be used in the SELECT list, between the SELECT and FROM keywords of a query.

<Continued>

## Defining window specifications—OVER clause

### Defining window specifications—OVER clause

When defining the window function:

- Include an OVER() clause.
  - Specify the window of data to which the function applies.
- Define:
- Window partitions, using the PARTITION BY clause.
  - Ordering within a window partition, using the ORDER BY clause.
  - Framing within a window partition, using ROWS and RANGE clauses.
  - The ORDER BY clause also defines a frame of unbounded preceding to current in the partition.

41

SQL Fundamentals

Dell EMC

All window functions must have an OVER() clause. The window function specifies the window of data to which the function applies.

It defines:

- Window partitions using the PARTITION BY clause.
- Ordering within a window partition using the ORDER BY clause.
- Framing within a window partition, for ROWS/RANGE clauses.
- The ORDER BY clause, which defines a frame of unbounded preceding to current in the partition.

## RANK and ORDER BY

### RANK and ORDER BY

The ORDER BY clause:

- Can always be used by window functions.
- Is required by some window functions, such as RANK.
- Specifies ordering within a window partition.

The RANK built-in function:

- Calculates the rank of a row.
- Gives rows with equal values for the specified criteria the same rank.

417

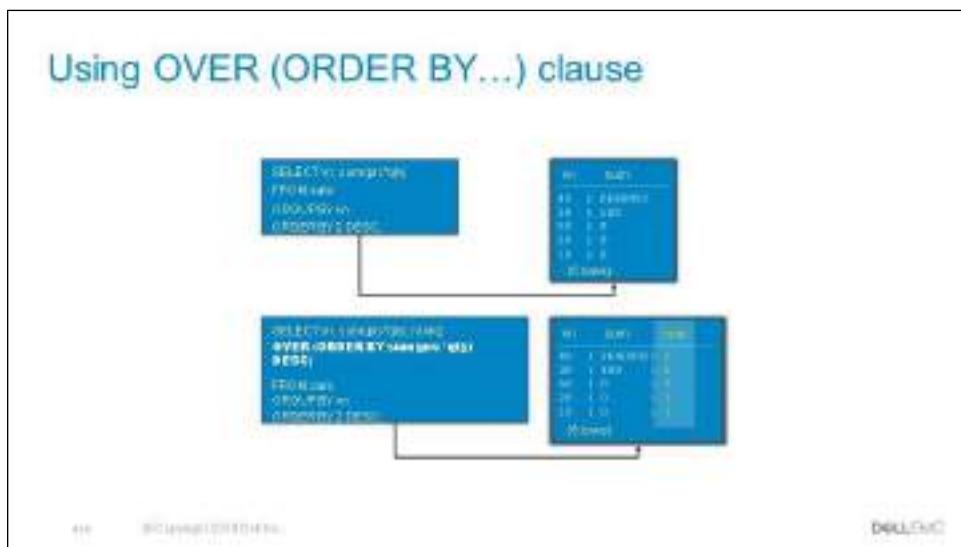
BIG DATA FOUNDATION

DELL EMC

The ORDER BY clause is used to order the resulting dataset based on an expression or column. It is always allowed in window functions and is required by some window functions, including RANK. The ORDER BY clause specifies ordering within a window partition.

The RANK function is a built-in function that calculates the rank of a row in an ordered group of values. Rows with equal values for the ranking criteria receive the same rank. The number of tied rows are added to the rank number to calculate the next rank value. In this case, ranks may not be consecutive numbers.

## Using OVER (ORDER BY...) clause



## Using the OVER (ORDER BY...) clause

The slide shows an example of two queries that rank vendors by sales totals. The first query shows a window function grouped on the vendor column, `vn`. The second query uses the `RANK` function to output a ranking number for each row. Note that the `PARTITION BY` clause is not used in this query. The entire result is one window partition. Also, do not confuse the `ORDER BY` of a window specification with the `ORDER BY` of a query.

## Window framing example

Window framing example

A rolling window moves through a partition of data, one row at a time.

SQL Statement:

```
SELECT vendor_id, date_id, sales, avg_sales
FROM fact_sales
OVER (PARTITION BY vendor_id
      ORDER BY date_id
      ROWS BETWEEN 1 PRECEDING AND 1 FOLLOWING)
```

DATASET

| Date ID | Sales | Avg Sales |
|---------|-------|-----------|
| 1       | 100   | 100       |
| 2       | 150   | 125       |
| 3       | 200   | 175       |
| 4       | 250   | 225       |
| 5       | 300   | 275       |
| 6       | 350   | 325       |
| 7       | 400   | 375       |
| 8       | 450   | 425       |
| 9       | 500   | 475       |
| 10      | 550   | 525       |
| 11      | 600   | 575       |
| 12      | 650   | 625       |
| 13      | 700   | 675       |
| 14      | 750   | 725       |
| 15      | 800   | 775       |
| 16      | 850   | 825       |
| 17      | 900   | 875       |
| 18      | 950   | 925       |
| 19      | 1000  | 975       |
| 20      | 1050  | 1025      |
| 21      | 1100  | 1075      |
| 22      | 1150  | 1125      |
| 23      | 1200  | 1175      |
| 24      | 1250  | 1225      |
| 25      | 1300  | 1275      |
| 26      | 1350  | 1325      |
| 27      | 1400  | 1375      |
| 28      | 1450  | 1425      |
| 29      | 1500  | 1475      |
| 30      | 1550  | 1525      |
| 31      | 1600  | 1575      |
| 32      | 1650  | 1625      |
| 33      | 1700  | 1675      |
| 34      | 1750  | 1725      |
| 35      | 1800  | 1775      |
| 36      | 1850  | 1825      |
| 37      | 1900  | 1875      |
| 38      | 1950  | 1925      |
| 39      | 2000  | 1975      |
| 40      | 2050  | 2025      |
| 41      | 2100  | 2075      |
| 42      | 2150  | 2125      |
| 43      | 2200  | 2175      |
| 44      | 2250  | 2225      |
| 45      | 2300  | 2275      |
| 46      | 2350  | 2325      |
| 47      | 2400  | 2375      |
| 48      | 2450  | 2425      |
| 49      | 2500  | 2475      |
| 50      | 2550  | 2525      |
| 51      | 2600  | 2575      |
| 52      | 2650  | 2625      |
| 53      | 2700  | 2675      |
| 54      | 2750  | 2725      |
| 55      | 2800  | 2775      |
| 56      | 2850  | 2825      |
| 57      | 2900  | 2875      |
| 58      | 2950  | 2925      |
| 59      | 3000  | 2975      |
| 60      | 3050  | 3025      |
| 61      | 3100  | 3075      |
| 62      | 3150  | 3125      |
| 63      | 3200  | 3175      |
| 64      | 3250  | 3225      |
| 65      | 3300  | 3275      |
| 66      | 3350  | 3325      |
| 67      | 3400  | 3375      |
| 68      | 3450  | 3425      |
| 69      | 3500  | 3475      |
| 70      | 3550  | 3525      |
| 71      | 3600  | 3575      |
| 72      | 3650  | 3625      |
| 73      | 3700  | 3675      |
| 74      | 3750  | 3725      |
| 75      | 3800  | 3775      |
| 76      | 3850  | 3825      |
| 77      | 3900  | 3875      |
| 78      | 3950  | 3925      |
| 79      | 4000  | 3975      |
| 80      | 4050  | 4025      |
| 81      | 4100  | 4075      |
| 82      | 4150  | 4125      |
| 83      | 4200  | 4175      |
| 84      | 4250  | 4225      |
| 85      | 4300  | 4275      |
| 86      | 4350  | 4325      |
| 87      | 4400  | 4375      |
| 88      | 4450  | 4425      |
| 89      | 4500  | 4475      |
| 90      | 4550  | 4525      |
| 91      | 4600  | 4575      |
| 92      | 4650  | 4625      |
| 93      | 4700  | 4675      |
| 94      | 4750  | 4725      |
| 95      | 4800  | 4775      |
| 96      | 4850  | 4825      |
| 97      | 4900  | 4875      |
| 98      | 4950  | 4925      |
| 99      | 5000  | 4975      |
| 100     | 5050  | 5025      |

## Window framing example

While window framing clauses require an ORDER BY clause, not all window functions allow framing.

The ROWS and RANGE clauses specify a positional or logical rolling window that moves through a window partition of data.

In the example shown here, the rolling frame applies to its partition—in this case, vendor—and ordering within that partition, date.

The example shows positional framing using the ROWS BETWEEN clause, where the result is interpreted with respect to the CURRENT ROW position in the partition. The **focus of the window frame moves from row to row within its partition only.**

## Designating sliding window

### Designating sliding window

A moving window:

- Defines a set of rows in a window partition.
- Allows you to define the first row and last row.
- Uses the current row as the reference point.
- Can be expressed in rows with the ROWS clause.
- Can be expressed as a range with the RANGE clause.

44

SQL Window Functions

DELL EMC

A moving or rolling window defines a set of rows within a window partition. When you define a window frame, the window function is computed with respect to the contents of this moving frame, rather than against the fixed content of the entire window partition. Window frames can be row based, represented by the ROWS clause, or value based, represented by a RANGE.

When the window frame is row based, you define the number of rows offset from the current row. If the window frame is range based, you define the bounds of the window frame in terms of data values offset from the value in the current row.

If you specify only a starting row for the window, the current row is used as the last row in the window.

The window frame can be defined as:

- **UNBOUNDED or expression PRECEDING**—This clause defines the first row of the window, using the current row as a reference point. The starting row is expressed in terms of the number of rows preceding the current row. If you define a ROWS window frame as 5 PRECEDING, the window frame starts at the fifth row preceding the current row. If the definition is for a RANGE window frame, the window starts with the first row whose ordering column value precedes that of the current row by 5. If the term UNBOUNDED is used, the first row of the partition acts as the first row of the window.

## Designating sliding window, cont.

### Designating sliding window, cont.

A moving window:

- Defines a set of rows in a window partition.
- Allows you to define the first row and last row.
- Uses the current row as the reference point.
- Can be expressed in rows with the ROWS clause.
- Can be expressed as a range with the RANGE clause.

417

DATA MANIPULATION

DELL EMC

- **UNBOUNDED or expression FOLLOWING**—This clause defines the last rows of the window, using the current row as a reference point. Similar to PRECEDING, the last row is expressed in terms of the number of rows following the current row. Either an expression or the term UNBOUNDED can be used to identify the last rows. If UNBOUNDED is used, the last row in the window is the last row in the partition.
- **BETWEEN window\_frame\_bound AND window\_frame\_bound**—This clause defines the first and last rows of the window, using the current row as a reference point. The first and last rows are expressed in terms of the number of rows preceding and following the current row, respectively. You can use other window frame syntax to define the bound. For example, BETWEEN 5 PRECEDING AND 5 FOLLOWING defines a window frame where the previous 5 rows and the next 5 rows from the current row are included in the moving window.
- **CURRENT ROW**—This clause references the current row in the partition. If a window frame is defined as BETWEEN CURRENT ROW AND 5 FOLLOWING, the window is defined starting with the current row and ending with the next 5 rows.

## Window framing example

The slide title is "Window framing example". Below it, a text box states: "A rolling window moves through a partition of data, one row at a time." A code snippet follows:

```
SELECT * FROM vendor_order_detail
ORDER BY vendor_id
ROWS BETWEEN CURRENT ROW AND CURRENT ROW + 2;
```

A table to the right shows data from January 1st to January 11th, 2010, with three columns: date, vendor\_id, and quantity.

| Date       | Vendor ID | Quantity |
|------------|-----------|----------|
| 2010-01-01 | 1000      | 3000     |
| 2010-01-01 | 2000      | 3000     |
| 2010-01-02 | 2000      | 5000     |
| 2010-01-03 | 2000      | 7000     |
| 2010-01-05 | 1000      | 1000     |
| 2010-01-06 | 2000      | 7000     |
| 2010-01-07 | 1000      | 7000     |
| 2010-01-08 | 2000      | 7000     |
| 2010-01-09 | 1000      | 7000     |
| 2010-01-10 | 2000      | 7000     |
| 2010-01-11 | 2000      | 7000     |

### Window framing example

While window framing clauses require an ORDER BY clause, not all window functions allow framing.

The ROWS and RANGE clauses specify a positional or logical rolling window that moves through a window partition of data.

In the example shown here, the rolling frame applies to its partition—in this case, vendor—and ordering within that partition, date.

The example shows positional framing using the ROWS BETWEEN clause, where the result is interpreted with respect to the CURRENT ROW position in the partition. The **focus of the window frame moves from row to row within its partition only.**

## General syntax of window function

### General syntax of window function

A moving window is defined as part of a window with the ORDER BY clause as follows:

```
WINDOW window_name AS (window_specification)
  where window_specification can be:
    [window_name]
    [PARTITION BY expression [, ...]]
    [ORDER BY expression {ASC|DESC|USING operator} [, ...]]
    [ROWS | RANGE]
    [UNBOUNDED PRECEDING
     | expression PRECEDING
     | CURRENT ROW
     | BETWEEN window_name_bound AND window_name_bound]]
```

Syntax of a moving window is specified here.

## Built-in window functions

| Built-in window functions    |  |
|------------------------------|--|
| Built-In Function            | Description  |
| dist()                       | Calculates the cumulative distribution of a value in a group of values. Rows with equal values always evaluate to the same cumulative distribution value.  |
| dense_rank()                 | Computes the rank of a row in an ordered group of rows without skipping rank values. Rows with equal values are given the same rank value.   |
| first_value(expr)            | Returns the first value in an ordered set of values.   |
| lag(expr [offset] [default]) | Provides access to more than one row of the same table without doing a self join. Given a series of rows returned from a query and a position of the cursor, LAG provides access to a row at a given physical offset before that position. If offset is not specified, the default offset is 1. default sets the value that is returned if the offset goes beyond the scope of the window. If default is not specified, the default value is null. |

Note: Any aggregate function used with the OVER clause can also be used as a window function.

## Built-in window functions

The slide shows built-in window functions supported within Greenplum. These built-in functions require an OVER clause.

For more detailed information about the functions, see the *Greenplum Database Administrator Guide*.

## Built-in window functions, cont.

### Built-in window functions, cont.

| Built-In Function | Description  |
|-------------------|--|
| last_value(expr)  | Returns the last value in an ordered set of values.  |
| lead()            | Provides access to more than one row of the same table without doing a self join. Given a series of rows returned from a query and a position of the cursor, LEAD provides access to a row at a given physical offset after that position. If offset is not specified, the default offset is 1. default sets the value that is returned if the offset goes beyond the scope of the window. If default is not specified, the default value is null. |
| ntile(expr)       | Divides an ordered dataset into various buckets—n, defined by expr—and assigns a bucket number to each row.  |
| percent_rank()    | Calculates the rank of a hypothetical row R minus 1, divided by 1 less than the number of rows being evaluated within a window partition.  |
| row_number()      | Assigns a unique number to each row to which it is applied—either each row in a window partition or each row of the query.   |

## Check your knowledge

### Check your knowledge

Describe how this code works:

```
SELECT dt, region, revenue,
       count(*) OVER (twdw) AS moving_count,
       avg(revenue) OVER (twdw) AS moving_average
  FROM moving_average_data mad
 WINDOW twdw AS (PARTITION BY region
 ORDER BY dt RANGE BETWEEN
 '7 days'::interval PRECEDING AND
 '0 days'::interval FOLLOWING)
 ORDER BY region, dt
```

40

SQL Fundamentals

Dell EMC

## Check your knowledge



### Discussion

## Question / Discussion Topic:

Take a look at the code shown here. How does this code work?

```
SELECT dt, region, revenue,
       count(*) OVER (twdw) AS moving_count,
       avg(revenue) OVER (twdw) AS moving_average
  FROM moving_average_data mad
 WINDOW twdw AS (PARTITION BY region
 ORDER BY dt RANGE BETWEEN
 '7 days'::interval PRECEDING AND
 '0 days'::interval FOLLOWING)
 ORDER BY region, dt
```

Here is how this code works:

- It will generate a table with five columns.
- The first three columns will be dt, region, and revenue ordered by region and dt.
- The next column will be a moving count, which will be a count of the number of values added to get the moving average.
- The last column will contain the average of revenue of the maximum of the last 7 days, based on the information available.

**Discussion Notes:**

## User-defined functions and aggregates

### User-defined functions and aggregates

Greenplum supports several function types, including:

- Query language functions where the functions are written in SQL.
- Procedural language functions where the functions are written in:
  - PL/pgSQL
  - PL/Tcl
  - Perl
  - Python
  - R
- Internal functions
- C-language functions
- Use case examples:
  - Second largest element in a column?
  - Online auction: Who is the second highest bidder?

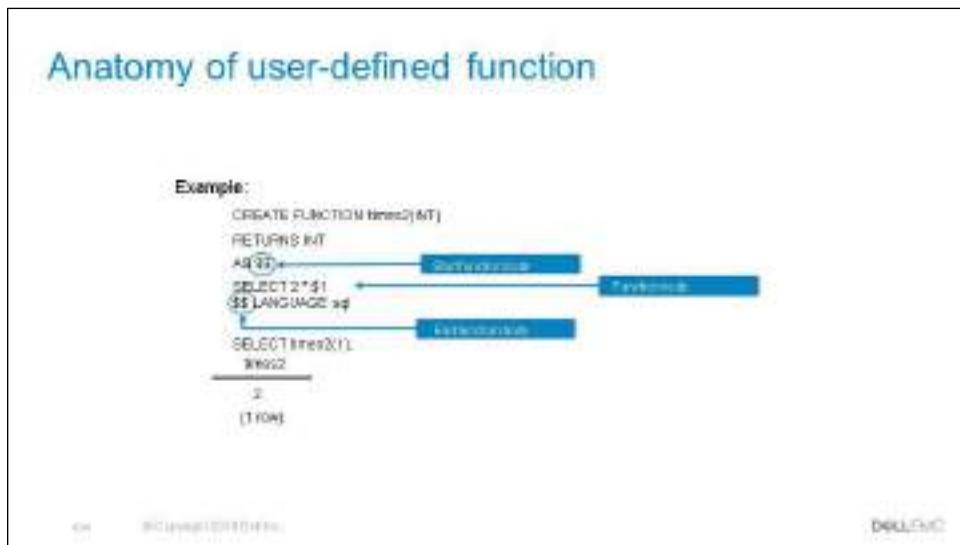
Greenplum supports a variety of methods for developing functions, including:

- Query language support for functions developed in SQL.
- Procedural language support for functions written in languages such as PL/PGSQL, which is a subset of PL/SQL, PL/Tcl, Perl, Python, and R, a programming language.
- Language for statistical computing and graphics.
- Internal functions.
- C-language functions.

The data scientist may need to create a function that could be used in the downstream analysis. Some use-case examples are shown here.

**Note:** Greenplum supports PL/pgSQL, PL/Perl, and PL/Python out of the box. Other languages can be added with the createlang utility.

## Anatomy of user-defined function



Here is a simple function that you can create.

Whenever you pass in a parameter, you can identify it as a base or primitive type, such as integer, char, or varchar.

In this example, when you call this function, you pass the parameter INT. The value is multiplied by 2 and returned as numeric. \$1 indicates the first parameter.

### **Creating, modifying, and dropping functions**

Functions that operate on tables must be created in the same schema. If you modify a table, you must have access to a schema. You:

- Create a function with the CREATE FUNCTION command. You must have CREATE access to the schema, to create a function. A function can be created with or without parameters.
- Replace an existing function with the CREATE OR REPLACE FUNCTION command. This command either creates a function, if one did not exist before, or replaces an existing function. If you are replacing an existing function, you must specify the same number of parameters and the same data types found in the original function. If not, you are creating a new function.

## Lesson: Advanced SQL and MADlib

- Change a function with the ALTER FUNCTION command. You must own the function before you can modify it. If the function is to be created in another schema, you must have CREATE privileges on that schema.
- Drop or remove a function with the DROP FUNCTION command. Because you can have multiple functions with the same name but different numbers of parameters and parameter types, you must include the appropriate number of parameters and parameter types as part of the command. You must also be the owner of the function, to remove the function from the schema.

## User-defined aggregates

**User-defined aggregates**

- Perform a single-table scan
- Example: Second-largest number
  - Keep a state: maximum two numbers
  - The new number can displace the smaller one in the state
  - Greenplum extension: Merge two states
- Example: Create a sum of cubes aggregate
 

```
CREATE FUNCTION cube_accum(aggregate, numeric) RETURNS numeric
AS '$1 + $2 * $2 * $2'
LANGUAGE SQL
IMMUTABLE
RETURNS NULL ON NULL INPUT;
CREATE AGGREGATE cube(numeric) (
  SFUNC = cube_accum,
  STYPE = numeric,
  INITCOND = 0);
```

User-defined aggregates perform a single table scan and it keeps state. A state is a maximum of two numbers.

This example shows the creation of a user-defined aggregate that returns a maximum of two numbers. You will learn more about this in the lab.

**CREATE AGGREGATE defines a new aggregate function.** Some basic and commonly used aggregate functions such as count, min, max, sum, avg, and so on, are already provided in the Greenplum database.

If one defines new types or needs an aggregate function not already provided, then CREATE AGGREGATE can be used to provide those features.

An aggregate function is made from one, two, or three ordinary functions—all of which must be IMMUTABLE functions. IMMUTABLE functions include a state transition function sfunc, an optional preliminary segment-level calculation function prefunc, and an optional final calculation function ffunc. These are used as follows:

- sfunc( internal-state, next-data-values ) □ next-internal-state
- prefunc( internal-state, internal-state ) □ next-internal-state
- ffunc( internal-state ) □ aggregate-value

In the example shown here, you only have the sfunc.

To test this aggregate, you can try the following code:

```
CREATE TABLE x(a INT);
```

## Lesson: Advanced SQL and MADlib

```
INSERT INTO x VALUES (1),(2),(3);
```

```
SELECT scube(a) FROM x;
```

Correct answer for reference:

```
SELECT sum(a*a*a) FROM x;
```

## Ordered aggregates

### Ordered aggregates

- Output of aggregates may depend on order:
  - Example:
    - `SELECT array_agg(letter) FROM alphabet`
  - SQL does not guarantee a particular order
  - Output could be (a,b,c) or (b,c,d) or ... depending on query optimizer, distribution of data, ...
- Sample use case:
  - Maximum value of discrete derivative? For example:
    - Largest single-day stock increase during last year?
- Greenplum introduces ordered aggregates.
  - `SELECT array_agg(column ORDER BY expression [ASC|DESC]) FROM table`
- Median can be implemented using an ordered call of `array_agg()`
  - This information is covered in the lab.

4/10 80% Complete

Dell EMC

Support has been added for ordered aggregate functions in Greenplum, providing a method for controlling the order in which values are fed to an aggregate function.

In a Greenplum database, only aggregate functions defined as ORDERED can be called with an ORDER BY clause. This can be followed by other arguments to specify a system-defined ordering.

The three built-in ordered aggregates and optional ORDER BY clauses that have been implemented in 4.1, are shown in the following table:

| Aggregate function                  | Description                             | Example   |
|-------------------------------------|---|---|
| <code>array_agg(any element)</code> | Concatenates any element into an array. | <code>SELECT array_agg(anyelement ORDER BY anyelement) FROM table;</code> |
| <code>string_agg(text)</code>       | Concatenates text into a string.        | <code>SELECT string_agg(text ORDER BY text) FROM table;</code>            |

## Lesson: Advanced SQL and MADlib

|                                |  |   |
|--------------------------------|--|---|
| string_agg(text,<br>delimiter) | Concatenates text<br>into a string<br>delimited by<br>delimiter. | SELECT string_agg(text, ',' ORDER<br>BY text) FROM table; |
|--------------------------------|--|---|

The columns in an ORDER BY clause are not necessarily the same as the aggregated column, as shown in the following statement that references a table named product with columns store\_id, product\_name, and quantity.

```
SELECT store_id, array_agg(product_name ORDER BY quantity desc) FROM  
product GROUP BY store_id;
```

**Note: There can only be one aggregated column. Multiple columns can be specified in the ORDER BY clause.**

## MADlib—definition

### MADlib—definition

- MAD stands for: Magnetic, Agile, Deep
- lib stands for library of:
  - Advanced (mathematical, statistical, machine learning)
  - Parallel and scalable
  - In-database functions
- Mission: to foster widespread development of scalable analytic skills, by harnessing efforts from commercial practice, academic research, and open-source development

4/17

© 2018 Dell Inc. All rights reserved.

DELL.COM

MADlib was first reported at VLDB 2009, in which **MAD Skills: New Analysis Practices for Big Data** was presented

- [db.cs.berkeley.edu/papers/vldb09-madskills.pdf](http://db.cs.berkeley.edu/papers/vldb09-madskills.pdf)

## MADlib in-database analytical functions

### MADlib in-database analytical functions

- Classification
- Regression
- Clustering and topic modeling
- Association rule mining
- Descriptive statistics
- Validation

40

Big Data Analytics v2

Dell EMC

Listed are examples of the in-database analytic functions available as MADlib functions. This list keeps changing with every update and as the user community contributes to the MADlib.

Visit [madlib.apache.org](http://madlib.apache.org) for the most recent version.

## Calling MADlib functions—fast training, scoring

### Calling MADlib functions—fast training, scoring

- MADlib allows users to easily and create models without moving data out of the systems.
  - Model generation
  - Model validation
  - Scoring—evaluation of—new data
- All the data can be used in one model.
- Built-in functionality to create of multiple smaller models—for example, regression/classification grouped by feature

Diagram illustrating the MADlib function call process:

- MADlib model location
- Table: training data
- Table: in which to write results
- Labels corresponding dependent variable
- MADlib function:
  - SELECT \* FROM train\_data
  - Table: training data
  - MADlib linear model
  - Table: in which to write results
  - Labels corresponding dependent variable
- Tables needed in this model
- Create model script based on this fit type

MADlib functions run directly on data in the HDFS system (using HAWQ). So, you do not have to retrieve data for analytical modeling purposes.

Example—Function to run linear regression model using MADlib.

Open-source lets you tweak and extend methods, or build your own.

## MADlib—getting help



## Lesson—summary

### Lesson—summary

During this lesson, the following advanced functions were covered:

- Window functions
- User-defined functions and aggregates
- Ordered aggregates
- MADlib



4.14      © Copyright 2018 Dell Inc.

DELL EMC

Advanced SQL and MADlib functions were covered in this lesson.

## Check your knowledge

### Check your knowledge

1. What is the difference between the GROUP BY and the Window function?
2. Can you create user-defined functions with multiple output parameters?
3. Give an example for Ordered Aggregates.
4. What are advantages of MADlib?

4/2

DATA SCIENCE

DELL EMC

## Check your knowledge



### Discussion

## Question / Discussion Topic:

1. What is the difference between the GROUP BY and the Window function?
2. Can you create user-defined functions with multiple output parameters?
3. Give an example for Ordered Aggregates.
4. What are advantages of MADlib?

## Discussion Notes:

## Module summary—advanced analytics—technology and tools

### Module summary—advanced analytics—technology and tools

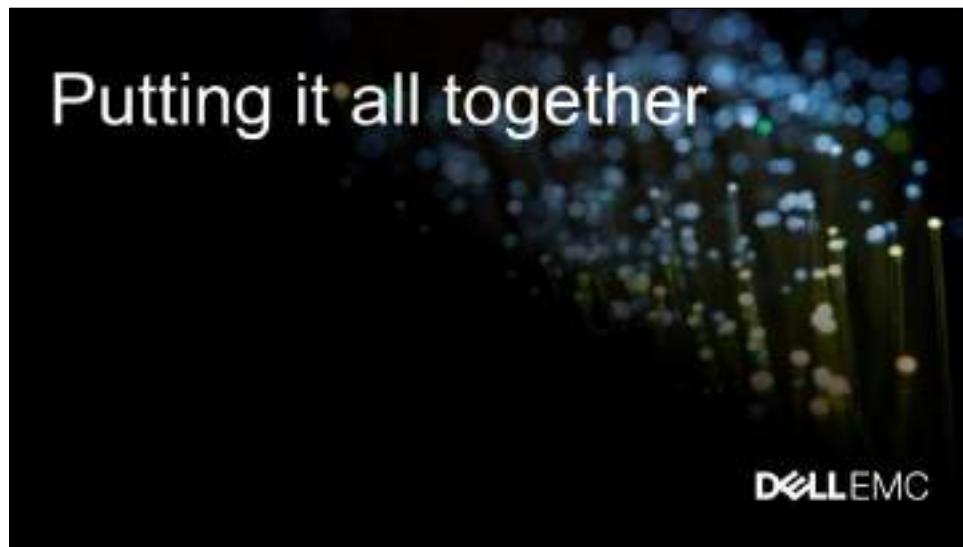
The key points covered in this module were:

- MapReduce, Hadoop, and the Hadoop ecosystems
- In-database analytics with advanced SQL functions and MADlib

These are the key topics covered in this module.

# Putting it all together

## Introduction



### Putting it all together

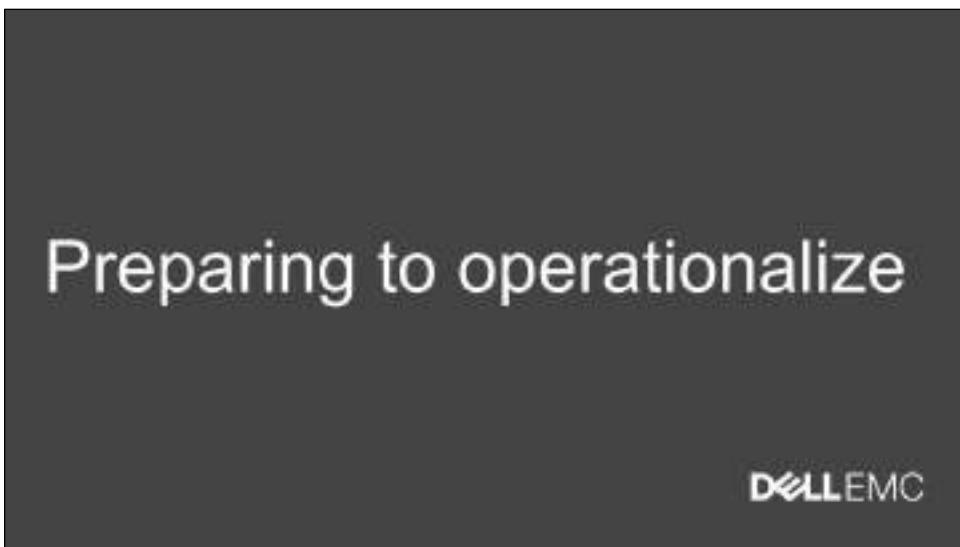
Upon completing this module, you should be able to:

- ✓ Articulate the tasks needed to operationalize an analytics project.
- ✓ Explain how four common project deliverables meet key stakeholders' needs.
- ✓ Prepare final presentations for sponsors and analysts.
- ✓ Evaluate a data visualization and identify ways to improve it.

This module focuses on communicating the results of an analytics project and considerations for implementing the model into production. This module also describes the use of a framework for creating final presentations for business sponsors and technical audiences. Further, this module covers the use of data visualizations for proper communication.

## Lesson: Preparing to operationalize

### Introduction



The slide has a dark gray background. In the center, the text "Preparing to operationalize" is displayed in a large, white, sans-serif font. In the bottom right corner, there is a small white Dell EMC logo.



The slide has a light gray background. At the top, the title "Preparing to operationalize" is shown in a teal-colored font. Below the title, the text "This lesson covers:" is followed by a bulleted list:

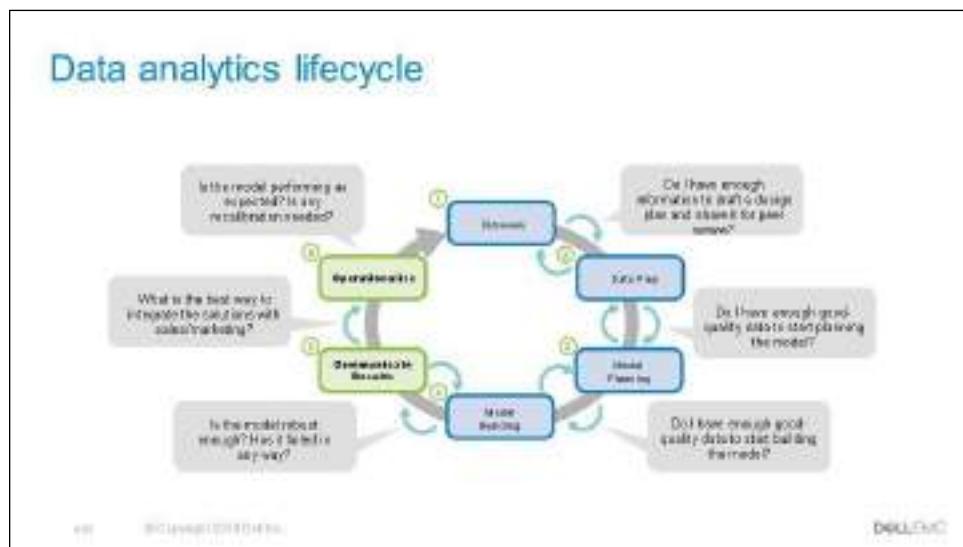
- Final phases of the data analytics lifecycle.
- Four core deliverables for a successful analytic project.

In the bottom left corner, there is a small white Dell EMC logo. In the bottom right corner, there is another small white Dell EMC logo.

This lesson covers important considerations for the successful of the final two phases of the data analytics lifecycle: communicate results and operationalize.

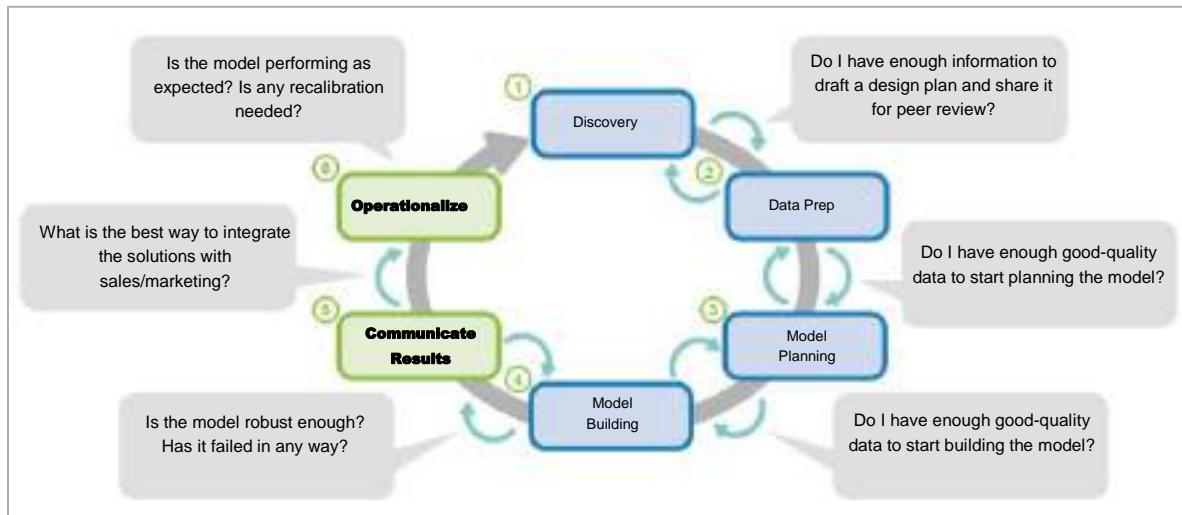
## Lesson: Preparing to operationalize

### Data analytics lifecycle



This graphic portrays the data analytics lifecycle. Through the first four phases:

- The business and analytic problems are defined.
- Data is explored and prepared for analysis.
- The analytic approach is developed and executed.



In the last two phases, the results of the analysis must be communicated to the various stakeholders and, if the business benefit is justified, the resulting model must be operationalized – that is, implemented into production.

## Typical end-of-project scenario

### Typical end-of-project scenario

- By the end of the model building phase:
  - The project team worked long and hard on gathering, cleaning, and munging data.
  - The "best model ever" is developed.
  - The team is proud of what was accomplished and the project findings.
- The project team now faces the following challenges:
  - Communicating the results to the various stakeholders outside of the project team
  - Convincing the decision makers to put the team's efforts into production
    - o Demonstrating the business value
    - o Addressing the concerns of critics or detractors
- Assuming the approval is granted, the model must be placed into production:
  - Will the same project team members implement the model?
  - Or, should new resources be brought up to speed?

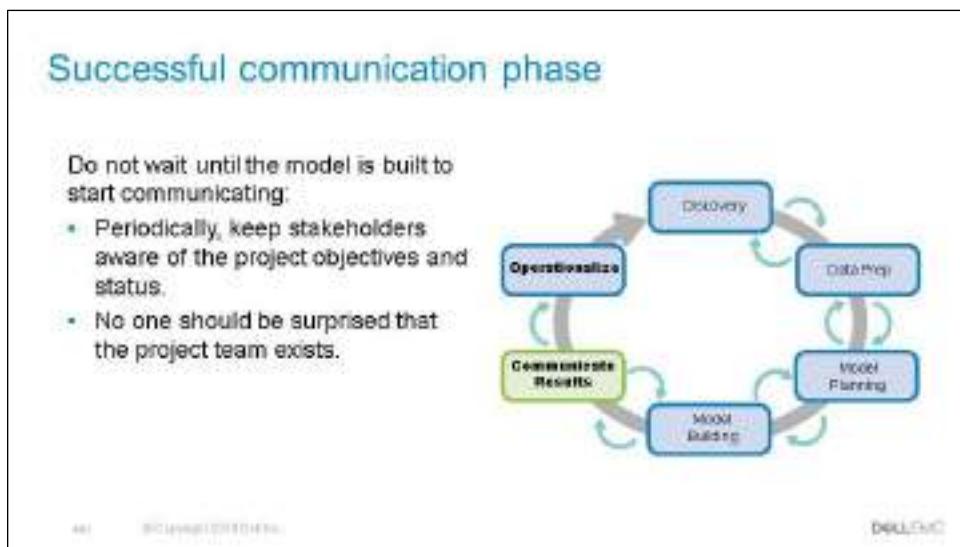
40

DATA SCIENCE

DELL INC.

The project team has worked long and hard on analyzing the data and building a model. The challenge now is how to effectively communicate the team's findings to the various audiences and stakeholders as well as to get approval to implement the model into production. Then, once approval is given to operationalize the model, there is the need to bring any new resources up to speed.

## Successful communication phase



Every organization is structured a little differently, and within an organization any project may run a little differently. There are various levels of integration between the lines of business, IT, finance, and other key functions. However, there are common themes that arise when communicating the business value of any proposal.

## Successful communication phase (cont. 1)

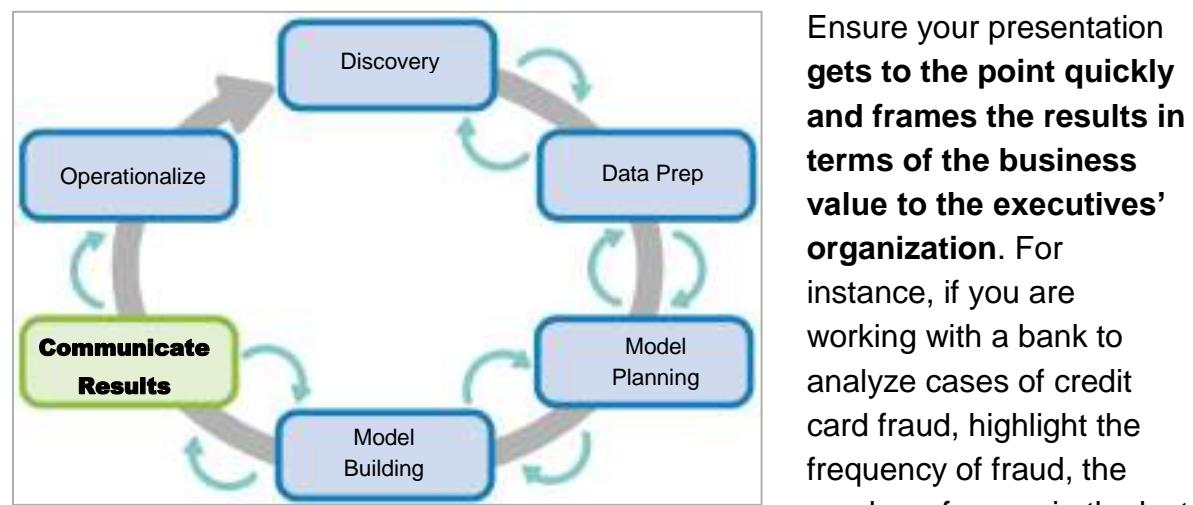
**Successful communication phase (cont. 1)**

Tailor the communications to the audience.

- Executives and business leaders care about business outcomes, not anyone's technical prowess.
  - Apply the experience of others who have presented to the same executives.
  - Obtain feedback and buy-in from the executives' staff.
- Business users care about how their process is impacted: favorably or negatively.
- Technical users need detailed information about data sources, code, and requirements.

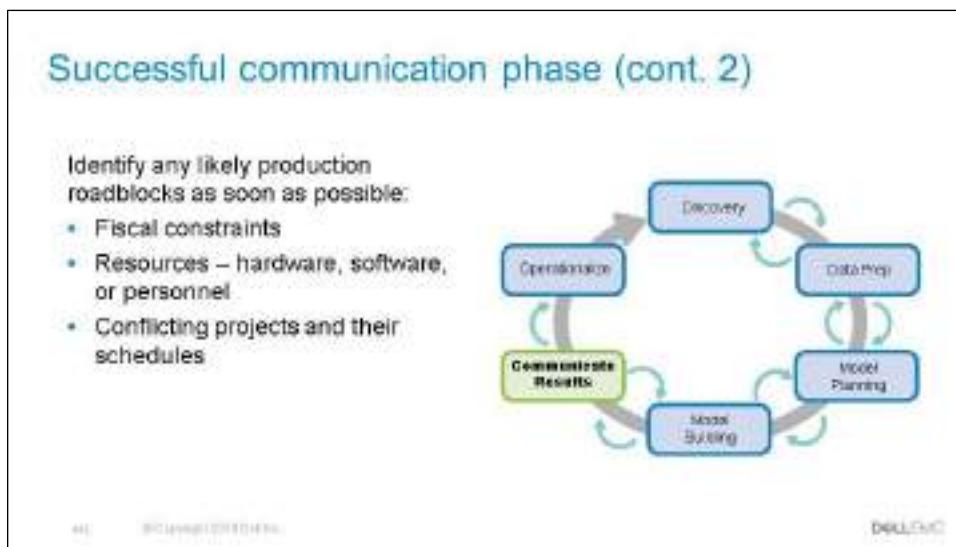
The diagram illustrates a cyclical process for successful communication. It consists of six blue rectangular boxes arranged in a circle, connected by curved arrows indicating a clockwise flow. The boxes are labeled: 'Discovery' (top), 'Data Prep' (top-right), 'Model Planning' (right), 'Model Building' (bottom-right), 'Operationalize' (bottom-left), and 'Communicate Results' (center). The 'Communicate Results' box is highlighted with a green border.

The more the audience is comprised of executives, the more succinct you must be. Most executive stakeholders attend many briefings in the course of a day or a week.



much of a cost or revenue impact there is to the bank. Or, focus on the reverse: how much more revenue they could gain if they address the fraud problem. This approach will demonstrate the business impact better than deep dives on the methodology.

## Successful communication phase (cont. 2)



You must include supporting information about analytical methodology and data sources, but generally only as supporting detail or to ensure the audience has confidence in the approach you took to analyze the data.

If presenting to analysts or other technical audiences, focus more time on the **methodology and findings**. You can afford to be more expansive in describing the outcomes, the methodology, and the analytical experiment with a peer group, because they will be more interested in the techniques, especially if you developed a new way of processing or analyzing data that can be reused in the future or applied to similar problems.

## Four core deliverables to meet stakeholders' needs

Four core deliverables to meet stakeholders' needs

- 1. Presentation for project sponsors and other executives
- 2. Presentation for analysts
- 3. Code
- 4. Technical specifications for implementing the code



DATA SCIENCE

These four items are key deliverables in any successful analytic project. Let's look at each of them in a little more detail.

## Four core deliverables to meet stakeholders' needs, cont.

### Four core deliverables to meet stakeholders' needs, cont.

1. Presentation for project sponsors and other executives
  - a. Identify "big picture" takeaways for executive-level stakeholders.
  - b. Determine key messages that can aid their decision-making process.
  - c. Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp.
  - d. Objective: Demonstrate the business benefit of implementing the model.
2. Presentation for analysts
  - a. Identify business process changes.
  - b. Discuss reporting changes.
  - c. Fellow data scientists want details and are comfortable with technical graphs; for example, ROC curves.
  - d. Objective: Demonstrate the validity of the model.
3. Code
4. Technical specifications for implementing the code

440

DATA SCIENCE

DELL INC.

Although there may be several key stakeholders – such as executives, data engineers, and business users – of an analytics project, most of their roles' responsibilities usually overlap at this stage of the project and can be met with four main deliverables. The two presentations will be instrumental in obtaining the business buy-in and approval to proceed with the operationalize phase. Two sets of technical documentation – code and specifications – will be necessary to begin the operationalize phase. These four items will be key deliverables in any successful analytic project.

Next, we'll focus briefly on the two technical deliverables; we'll cover the presentations in another lesson.

## Considerations for technical documentation and code

### Considerations for technical documentation and code

- Provide well-documented code with meaningful variable names.
- Describe the input data elements and sources:
  - Document the expected input data format and associated units.
  - Describe the preprocessing steps before data goes to the model code.
- Execute the model:
  - Explain the model processing.
  - Identify possible error situations—for example, missing data elements:
    - Provide guidance for dealing with any exception.
    - Specify any default conditions or outputs.
- Describe how to interpret and apply the model outputs.
- Address the need to easily update the model. Identify what will likely change over time—for example, model coefficients or variables.

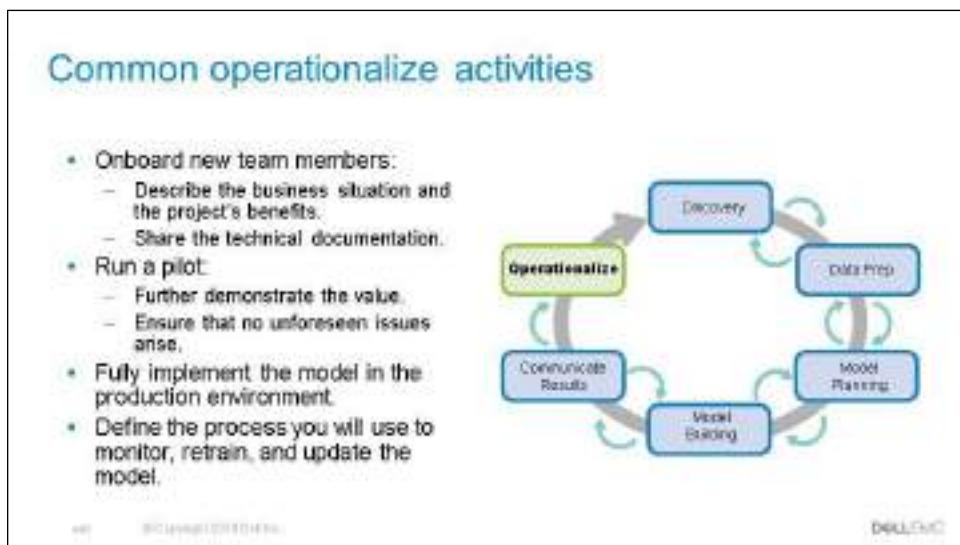
44

DATA SCIENCE

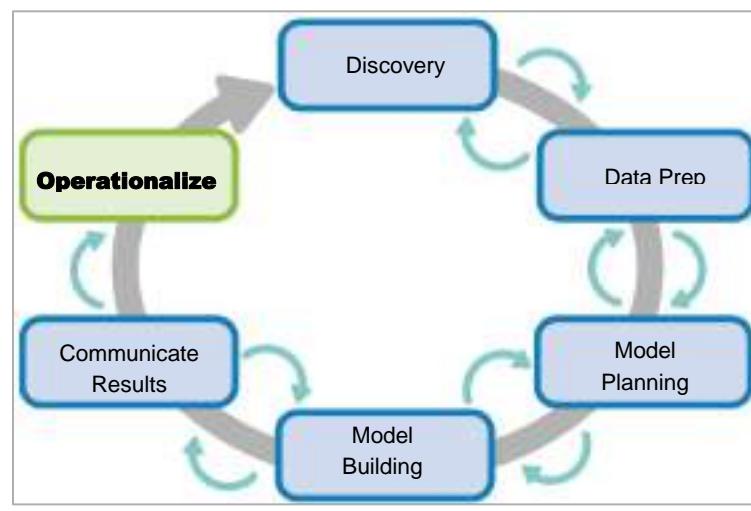
DELL INC.

Often, the code and the documentation will be given to a different team to place into production. So, it is important to provide readable code with meaningful variables names.

## Common operationalize activities



The operationalize phase focuses on implementing the results of the project. Unless the organization utilizes DevOps practices, it is often necessary to add new team members to the team or to hand-off the work to a different team. In either case, it is necessary to onboard the new resources and provide the technical requirements. Next, it is common to run a pilot program before fully implementing the model into production. Running a pilot will help minimize risk and further demonstrate the business value.



Consider running the model in a product environment for a discrete set of single products or a single line of business, to test your model in a live setting. This allows you to learn from the deployment and make any needed adjustments before launching across the enterprise.

Once the model is placed into production, it is often necessary to monitor the model's performance and establish a process to retrain and update the model.

## Lesson: Preparing to operationalize

It should be noted that further communication of results often occurs during the operationalize phase, so the executives can determine the actual ROI from their investment.

## Check your knowledge

**Check your knowledge**

Which stakeholder of an analytics project typically asks questions related to the business impact, the risks, and the return on investment of the project?

A. Business user      C. Project sponsor  
B. Project manager      D. Data scientist

44%      55% Unanswered      Dell EMC

### Check your knowledge:

Which stakeholder of an analytics project typically asks questions related to the business impact, the risks, and the return on investment of the project?

- A. Business user
- B. Project manager
- C. Project sponsor
- D. Data scientist

### Question:

Which stakeholder of an analytics project typically asks questions related to the business impact, the risks, and the return on investment of the project?

### Answer:

## Check your knowledge

**Check your knowledge**

What are the main deliverables provided to application developers and IT professionals who may manage the production environment?

A. Presentation for analysts      C. Code  
B. Presentation for project sponsor      D. Technical specifications

DELL EMC

### Check your knowledge:

What are the main deliverables provided to application developers and IT professionals who may manage the production environment?

- A. Presentation for analysts
- B. Presentation for project sponsor
- C. Code
- D. Technical specifications

### Question:

What are the main deliverables provided to application developers and IT professionals who may manage the production environment?

### Answer:

## Lesson: Preparing project presentations

### Introduction

Preparing project presentations

DELL EMC

### Preparing project presentations

This lesson covers:

- Using the analytic plan to guide presentation development
- Preparing the proper presentation material for the intended audience
- Documenting code and technical requirements

481 800x600x300px

DELL EMC

This lesson covers the final deliverables in detail for a case study SuperMom&PopShop scenario. The topics covered in this lesson include the analytic plan to guide the final presentation and its key aspects. Further, this lesson also covers how to develop and deliver the presentations. Finally, the lesson covers the overview of the code and technical documentation.

## Key aspects of project presentations

### Key aspects of project presentations

- Reflect on the project.
- Address the needs and concerns of the audience.
- Focus on the "how" for the analysts.



© 2018 Dell Inc. All rights reserved. Dell EMC

The three key aspects of creating project presentations are as follows:

- Reflect on the project.
- Address the needs and concerns of the audience.
- And, focus on the “how” for the analysts.

## Key aspects of project presentations (cont. 1)

### Key aspects of project presentations (cont. 1)

Reflect on the project:

- Consider the context of the problems you set out to solve.
- Identify observations about the model outputs, scoring, and results.
- Identify key messages and any unexpected insights.



Ideally, the team should consider starting the development of the final presentation during the project, rather than at the very end of the project. This approach ensures that the team always has a version of the presentation with working hypotheses to show stakeholders in case there is a need to show a work-in-process (WIP) version of the project progress on short notice.



In fact, many analysts write the executive summary at the outset of a project, and then continually refine it over time so that, at the end of the project, portions of the final presentation are already completed. This approach also reduces the chance that the team members would forget key points or insights discovered during the project. Finally, it reduces the amount of work to be

done on the presentation at the end of the project.

## Key aspects of project presentations (cont. 2)

### Key aspects of project presentations (cont. 2)

Address the needs and concerns of the audience.

Focus on the "what" for the sponsors:

- Show that the project goals were met.
- Provide the business benefits of implementing the model.
- Provide talking points for the executive to evangelize the project.

The project sponsor presentation focuses on the “what” aspects of the project. For the project sponsor, show that the team met the project goals. Focus on what was done, what the team accomplished, what return on investment (ROI) can be anticipated, and what business value can be realized. Give the project sponsor talking points to evangelize the work. Remember that the sponsor must convey the story to the others.

So, make this person’s job easy, and help ensure that the message is accurate by providing the few talking points. Find ways to emphasize ROI and business value, and mention whether the models can be deployed within performance constraints of the sponsor’s production environment.

## Key aspects of project presentations (cont. 3)

### Key aspects of project presentations (cont. 3)

Address the needs and concerns of the audience.

Focus on the "how" for the analysts:

- Show how the project goals were met.
- Explain how the model works and its dependencies – for example, data inputs.
- Discuss how to implement the model in products.
- Describe how current processes may be affected.

When presenting to a technical audience, such as data scientist and analysts, focus on how the work was done. Discuss how the team accomplished the goals and choices it made in selecting models or analyzing the data.

Also, share analytical methods and decision-making processes to the analysts. Describe methods, techniques, and technologies used, because this technical audience is interested in learning and considering whether it can be extended to other similar projects. Plan to provide specifics related to model accuracy and speed, such as how well the model performs in the production environment.

## Analytic plan for SuperMom&PopShop scenario

| Plan Component                                 | Details   |
|--|---|
| Business problem framed as an analytic problem | What is the likelihood – or, probability – of a customer churning?  |
| Initial hypotheses:                            | <ul style="list-style-type: none"> <li>▪ Purchase frequency and number of SKU's purchased are key predictors of churn.</li> <li>▪ Promotions and special offers can cost-effectively prevent churn.</li> </ul>  |
| Data and scope:                                | <p>For the last 5 years, obtain:</p> <ul style="list-style-type: none"> <li>▪ Customer demographics</li> <li>▪ Customer purchase details – in-store and online           <ul style="list-style-type: none"> <li>▫ For example: SKU / quantity / price per unit / discounts applied / type of purchase / purchase location / addresses</li> </ul> </li> <li>▪ Method of purchase: cash, debit, credit card, or store credit</li> <li>▪ Return details</li> </ul> |

Here is an analytic plan that has been updated through the Model Building phase for an aspect of the SuperMom&PopShop scenario where the company is looking to reduce customer churn. In addition to guiding your model planning and methodology, the analytic plan can serve as a guide for the main points of the final presentation.

Within the analytic plan are components that can be used as inputs for writing about the scope, underlying assumptions, modeling techniques, initial hypothesis, and key findings. After spending the time in modeling and performing in-depth data analysis, it is critical to reflect on the project work and consider the context of the problem the team set out to solve.

## Lesson: Preparing project presentations

An updated analytic plan can help focus the key points to be communicated to the sponsors and analysts.

### Analytic plan for SuperMom&PopShop scenario

| Plan Component                                  | Details   |
|---|---|
| Business problem framed as an analytics problem | What is the likelihood – or, probability – of a customer churning?  |
| Initial hypotheses                              | <ul style="list-style-type: none"><li>Purchase frequency and number of SKUs purchased are key predictors of churn.</li><li>Promotions and special offers can cost-effectively prevent churn.</li></ul>  |
| Data and scope                                  | For the last 5 years, obtain: <ul style="list-style-type: none"><li>Customer demographics</li><li>Customer purchase details – in-store and online<ul style="list-style-type: none"><li>For example: SKU / quantity / price per unit / discounts applied / type of purchase / purchase location / addresses</li><li>Method of purchase: cash, debit, credit card, or store credit</li><li>Return details</li></ul></li></ul> |

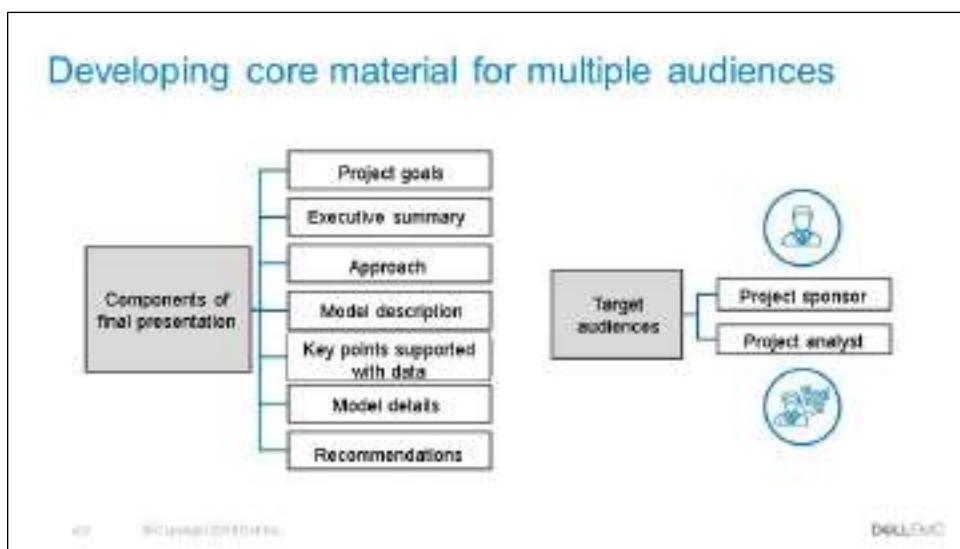
## Analytic plan for SuperMom&PopShop scenario (cont.)

| Plan Component (cont.)             | Details (cont.)  |
|------------------------------------|--|
| Model planning: analytic technique | Apply a logistic regression model: <ul style="list-style-type: none"><li>- To identify the most influential predictors of churn</li><li>- To estimate a probability of churn, and assign a churn number or churn label</li></ul>   |
| Results and key findings           | <ul style="list-style-type: none"><li>- What are the top predictors of customer churn?<ul style="list-style-type: none"><li>- Lack of an online account</li><li>- Credit card on file</li><li>- Distance from store</li></ul></li><li>- What is the most profitable approach to retain a customer?<ul style="list-style-type: none"><li>- Provide 30% discount offers after 60 days of inactivity</li><li>- For customers living greater than 30 miles away from a store, offer free online shipping</li></ul></li></ul> |
| Business Impact                    | This plan is projected to reduce customer attrition by 30 percent, avoiding \$2 million dollars in lost revenue.   |

## Analytic plan for SuperMom&PopShop scenario (cont.)

| Plan Component                     | Details  |
|------------------------------------|--|
| Model planning: analytic technique | <p>Apply a logistic regression model:</p> <ul style="list-style-type: none"><li>• To identify the most influential predictors of churn</li><li>• To estimate a probability of churn, and assign a churn number or churn label</li></ul>  |
| Results and key findings           | <ul style="list-style-type: none"><li>• What are the top predictors of customer churn?<ul style="list-style-type: none"><li>– Lack of an online account</li><li>– Credit card on file</li><li>– Distance from store</li></ul></li><li>• What is the most profitable approach to retain a customer?<ul style="list-style-type: none"><li>– Provide 30% discount offers after 60 days of inactivity.</li><li>– For customers living greater than 30 miles away from a store, offer free online shipping.</li></ul></li></ul> |
| Business impact                    | This plan is projected to reduce customer attrition by 30 percent, avoiding \$2 million dollars in lost revenue.   |

## Developing core material for multiple audiences



This graphic shows the main components of the final presentation for the project sponsor and for the analyst audiences. Notice that it can be helpful to create a core set of materials in these seven areas that can be used for two presentation audiences.

## Components of final presentation—project goal

### Components of final presentation—project goal

#### Situation synopsis – optional

- **Situation:** Revenue has dropped 15% over the last four quarters.
- **Complication:** The customer churn rate is 23 percent; with a 44 percent cut in the advertising budget.
- **Implication:** A continued need to cut expenses results in continued revenue losses.



The project goals portion of the final presentation is generally the same, or similar, for sponsors and for analysts. It may make sense to provide a summary of the business situation that was the impetus to initiate the project.

One method for writing the situational overview in a succinct way is to summarize it in the following three bullets.

- **Situation:** Give a succinct overview of the situation that has led to the analytical project.
- **Complication:** Give a one-sentence overview of the need for addressing this now. Something has triggered the organization to decide to take action at this time. Usually, this sentence represents



## Lesson: Preparing project presentations

the driver for why a particular project is being initiated at this point in time, rather than at some vague time in the future.

- **Implication:** Give a one-sentence overview of the impact of the complication. For instance, if the bank doesn't address their customer attrition problem, they stand to lose their dominant market position in three key markets. Focus on the business impact to illustrate the urgency of doing the project.

## Components of final presentation—project goal (cont.)

### Components of final presentation—project goal (cont.)

#### Project goals

- Predict the likelihood of an individual customer churning.
- The churn model's predictive power should be at least 10 percent better than the current technique.
- Churn rate should be reduced to 15 percent.

4/2

DATA SCIENCE & ANALYTICS

DELL EMC

Since the project goals portion of the final presentation is generally the same – or similar – for sponsors and for analysts, for each audience, the team must reiterate the goals of the project to lay the groundwork for the solution and recommendations that are shared later in the presentation.

In addition, the goals slide serves to ensure there is a shared understanding between the project team and the sponsors and to confirm that they are aligned in moving forward in the project. Generally, the goals are agreed on early in the project, and it is good practice to write them down and share them to ensure the goals are clearly understood by both the project team and the sponsors.

Project goals are provided for the SuperMom&PopShop scenario.

## Components of final presentation—executive summary

**Components of final presentation—project goal**

Executive summary

Incorporation of medical records in the electronic health record system (EHR) will reduce the cost of treatment by 10% to 15%.  
Health records of 100 patients are expected.

Key message to be offered in a 30-second duration after 60 days of inactivity.  
Distance from the doctor is a key predictor of health.  
A good doctor is trying to identify types of disease and understand conditions from the patient's narrative.

400 800 1200 1600 2000 2400 2800 3200 3600 4000 4400 4800 5200 5600 6000 6400 6800 7200 7600 8000 8400 8800 9200 9600 10000 10400 10800 11200 11600 12000 12400 12800 13200 13600 14000 14400 14800 15200 15600 16000 16400 16800 17200 17600 18000 18400 18800 19200 19600 20000 20400 20800 21200 21600 22000 22400 22800 23200 23600 24000 24400 24800 25200 25600 26000 26400 26800 27200 27600 28000 28400 28800 29200 29600 30000 30400 30800 31200 31600 32000 32400 32800 33200 33600 34000 34400 34800 35200 35600 36000 36400 36800 37200 37600 38000 38400 38800 39200 39600 40000 40400 40800 41200 41600 42000 42400 42800 43200 43600 44000 44400 44800 45200 45600 46000 46400 46800 47200 47600 48000 48400 48800 49200 49600 50000 50400 50800 51200 51600 52000 52400 52800 53200 53600 54000 54400 54800 55200 55600 56000 56400 56800 57200 57600 58000 58400 58800 59200 59600 60000 60400 60800 61200 61600 62000 62400 62800 63200 63600 64000 64400 64800 65200 65600 66000 66400 66800 67200 67600 68000 68400 68800 69200 69600 70000 70400 70800 71200 71600 72000 72400 72800 73200 73600 74000 74400 74800 75200 75600 76000 76400 76800 77200 77600 78000 78400 78800 79200 79600 80000 80400 80800 81200 81600 82000 82400 82800 83200 83600 84000 84400 84800 85200 85600 86000 86400 86800 87200 87600 88000 88400 88800 89200 89600 90000 90400 90800 91200 91600 92000 92400 92800 93200 93600 94000 94400 94800 95200 95600 96000 96400 96800 97200 97600 98000 98400 98800 99200 99600 100000

Writing a solid executive summary to portray the main findings of a project is crucial. In many cases, it may be the only portion of the presentation hurried managers read. For this reason, it is imperative to make the language clear, concise, and complete.



Someone consuming the executive summary should be able to grasp the full story of the project and the key insights in a single slide. In addition, this is an opportunity to provide key talking points for the executive sponsor to use to evangelize the project work with others in the customer's organization.

Be sure to frame the outcomes of the project in terms of business value, which is especially important if the presentation is for the sponsor.

We recommend making your key message very clear and conspicuous at the forefront of the slide. This can be set apart with color, as shown, or you can use some other technique to draw attention to it. Your key message may become the talking point that executives or the project sponsor takes away from the project and uses to support your recommendation for a pilot project, so this must be succinct.

## Lesson: Preparing project presentations

and compelling. To make this message as strong as possible, measure the value of the work and quantify the cost savings, revenue, time savings, or other benefits to make the business impact concrete.

Support the key message with about three major points. You can have more than three major points. However, going beyond three ideas makes it difficult for people to recall the main points; so, you must ensure that you keep the ideas clear and limit the summary to the few most impactful ideas you want the audience to take away from your work. If you list 10 key points, your message becomes diluted and the audience may remember only one or two.



In addition, since it is an analytical project, be sure to make one of your key points related to if, and how well, your work will meet the sponsor's SLA.

## Components of final presentation—executive summary (cont.)

Components of final presentation—executive summary (cont.)

Emphasize key messages.

Material Analysis

LBI Chart

Enables reader to grasp full project synopsis in one slide

Frames outcomes in terms of business value

Overall Response Rate = 3%

Dell EMC

Finally, although not required, it may be a good idea to support your main points with a visual, which will support the major points written at right. Visual imagery will serve to make a visceral connection with the readers, and help them retain the main message.

## Components of final presentation—approach

### Components of final presentation—approach



High-level methodology

- Interviewed 16 SuperMom&PopShop team members to understand current market conditions and business processes
- Collaborated with IT to identify relevant datasets and assess data quality
- Developed churn model to identify customers who will not purchase again in six months
  - Considered 43 input variables
  - Identified 11 most influential factors
  - 33 percent better predictive power than current model
- Prototyped model with IT to optimize performance in production

DATA SCIENCE

In the approach portion of the presentation, you need to explain the methodology you pursued on the project. This can include interviews with domain experts, cross-functional collaborations, and a few statements about the solution you developed. The objective of this slide is to ensure the audience is clear on the course of action you pursued and understands it well enough to explain it to others.

Also, be sure to include any additional comments related to your working assumptions as you did the work, this can be critical in defending why you pursued a specific course of action. When explaining the solution you developed, keep it at a high level for the project sponsors.

Finally, as part of the description on your approach, you may also want to mention constraints from systems, tools, or existing processes, and any implications for how these things may need to change due to this project.

## Components of final presentation—approach (cont.)

Components of final presentation—approach (cont.)



High-level methodology  
Relevant details on modeling techniques and technology

- Interviewed the members of the retailer to understand SuperMomsPopShop market and operating cost pressures for customer retention
- Collaborated with IT to identify relevant datasets and to assess data quality and availability
- Developed 'chain reader' in R using logistic regression
  - Three separate prediction models based on customer segments:
    - = On-store only
    - = Online only
    - = On-store and online
  - 30 percent better predictive power than current model, with 17 percent higher false positives
- Performed A/B testing to identify optimal discount offerings
- In the test environment, tested the SQL queries for use in production

DATA SCIENCE

Remember, if presenting to analysts or data scientists, add more detail about the type of model used, the technology, and the actual performance of the model during your tests.

## Components of final presentation—model description

Components of final presentation—model description

Overview of modeling technique

Overview of modeling technique

- Methodology overview
  - Identify the most predictive input variables.
  - Define any models that do not improve the model.
- Model: Logistic regression
- Data: Analysis
  - Point-of-sale system for in-store transactions.
  - MySQL database for online transactions.
  - CRM system for customer information and complaints.
  - Credit history reports.
- Scope of data
  - 1.1 billion transactions and 400,000 data points from Jan 1, 2013 to Dec 31, 2017 – Big work!
  - 683,000 unique customers.
  - 88.620% of customers are female and 11.379% are male.

Next: [Data Science and Big Data Analytics v2](#)

DELL EMC

Although the model description slide can be used for both audiences, the interests and objectives differ for each audience.

For the sponsor, articulate the general methodology without getting into too much detail. Convey the basic methodology followed in your work so the sponsor can communicate this to others within the organization.

To do this, focus on explaining the general methodology you used in a way that enables your sponsor to convey it to others, and provide talking points. Mention the scope of the data used to illustrate thoroughness and provide confidence that you used an approach that was an accurate portrayal of their problem and as free from bias as possible. One of the key traits of a good data scientist is the ability to be skeptical of one's own work. This is an opportunity to view the work and the deliverable with a critical eye and consider how it is received by the audience.

## Components of final presentation—key points with data

Components of final presentation—key points with data

The infographic is titled "Components of final presentation—key points with data". It features two main sections: "Support the key points with simple charts and graphics, such as bar charts." and "To support the key points, more details: Use analyst-oriented charts and graphs, such as ROC curves and histograms." Below these are two sample charts: "Predicted vs Actual Table 2" (dot density plot) and "ROC Curve Results" (ROC curve).

Support the key points with simple charts and graphics, such as bar charts.

To support the key points, more details: Use analyst-oriented charts and graphs, such as ROC curves and histograms.

Support the key points with charts and visualizations.

Predicted vs Actual Table 2

Actual = Predicted

Dot Density

Row

ROC Curve Results

AUC: 0.931

ROC Curve Results

Dell EMC

When developing key points, consider the insights that drive the biggest business impact and are defensible with data.

**For project sponsors**, use simple charts such as bar charts, which illustrate data clearly and allow the audience to understand the value of the insights. This is also where you foreshadow some of your recommendations and begin tying together your ideas to demonstrate what led to your recommendations and why. Creating clear, compelling slides to show your key points make the recommendations more credible and more likely to be acted upon by the customer.

**For analyst presentations**, you need more granular graphics. In this case, you may want to show a dot density chart or a histogram of a data distribution to support decisions you made in your modeling techniques. We will further discuss basic concepts of data visualizations in the next lesson.

## Components of final presentation—model details

### Components of final presentation—model details



Only include this icon if you are sharing the model with people who have no technical understanding.



Demonstrate the main logic or expression of the model.  
Provide any variables and associated coefficients.  
Describe technologies or tools used to execute the model.  
Describe expected model performance and any caveats.

400      01040001000000000000000000000000      DELL EMC

Model details are needed to share additional information with people who have a more technical understanding than the sponsors, those who need to implement the code, or colleagues on the analytical team.

## Components of final presentation—model details (cont.)

### Components of final presentation—model details (cont.)

Depending on its complexity, the model can be explained in one to dozens of slides:

Code snippets are fine, but do not do a code review in the presentation.

Revisit Code and detailed technical documentation file provided.

Focus on the details that best help the analyst understand the model and its use in production.

#### Possible discussion points include:

Input data cleaning and transformations.

Expected input formats.

Processing the inputs via the model logic.

Interpreting the model output.

Actions to take on the model output.

|                 |              | Churn        |       |
|-----------------|--------------|--------------|-------|
|                 |              | Actual Churn |       |
| Predicted Churn | Actual Churn | 0            | 1     |
|                 | 0            | 1980         | 307   |
| 1               | 1            | 918          | 10180 |

DELL EMC

In this section, discuss the variables used in the model, and explain how or why you selected the ones you did. In addition, share the essential code snippets to explain the basic mechanics. You can also use this section to illustrate details of the key variables and the predictive power of the model, using analyst-oriented charts and graphs, such as histograms, dot density charts, and ROC curves. Discuss the speed with which the model can run in the test environment, the expected performance in a live, production environment, and the technology needed.

This kind of discussion will address how well the model can meet the organization's SLA. Finally, include any additional caveats of the model and model performance, such as systems or data the model will need to interact with, performance issues, or how to feed the outputs of the model into existing business processes. Describe the relationships of the main variables such as the effects of key variables on predicting churn or the relationship of key variables to other variables.

## Components of final presentation—recommendation

### Components of final presentation—recommendation



Focus on business impact of the project, including risks and ROI.  
Give the sponsor key takeaways to help them evangelize the work within their organization.



Supplement recommendations with any implications for the modeling, or for deploying in a production environment.

Possible recommendations and next steps:

- Pilot the churn model in production with more A/B testing.
- Implement monitoring and reporting to ensure the model's continued effectiveness.
- Maintain the analytic sandbox for future analyses.
- Consider extra data sources to further improve the model:
  - Targeted customer surveys to investigate the causes of the churn
  - Social media

400DATA SCIENCE AND ANALYTICSDELL INC.

Create a set of recommendations that include how to deploy the model from a business perspective within the organization and any other suggestions on the rollout of this logic, depending on your knowledge of the domain area from the discovery phase. Attempt to measure the impact of the improvements and state how to use this within the recommendations.

By this point in the presentation, the preceding materials should have laid the groundwork for these recommendations.

Of course, certain audiences or organizations may prefer to see the content in a modified order. If the presentation is necessarily long, it may be wise to move the recommendations up early in the slide deck or at least address the main purpose of the presentation in the first few minutes – for example, “Today, the project findings will be presented with the intent of securing approval to pilot the churn model in production.”

## Summary of core components for target audiences

| Presentation component         | Project sponsor presentation  | Analyst presentation   |
|--------------------------------|---|--|
| Project goals                  | + List top 3 agreed-upon goals.   | + List top 3 agreed-upon goals.  |
| Executive summary              | + Emphasize key messages.   | + Emphasize key messages.  |
| Approach                       | + Discuss high-level methodology.   | + Discuss high-level methodology.<br>+ Provide relevant details on modeling techniques and technology.   |
| Model description              | + Provide overview of the modeling techniques.                                  | + Provide overview of the modeling techniques.   |
| Key points supported with data | + Support key points with simple charts, and graphics—for example, a bar chart. | + To support the key points, show details.<br>+ Use analyst-oriented charts and graphs, such as ROC curves and histograms.<br>+ Show visuals of key variables, and discuss significance of each. |

The table on this slide shows the summary of main components of the final presentations for the project sponsor and for the analyst audiences.

Three areas—such as project goals, executive summary, and model description – can be used as-is for both presentations. Other areas need more elaboration, such as the approach. Still other areas, such as the key points, require different levels of detail for the analysts and data scientists than for the project sponsor.

## Lesson: Preparing project presentations

| Presentation component         | Project sponsor presentation<br>* = Same components for both presentations  | Analyst presentation<br>* = Same components for both presentations   |
|--------------------------------|---|--|
| Project goals                  | <ul style="list-style-type: none"> <li>List top 3 agreed-upon goals.</li> </ul>   | <ul style="list-style-type: none"> <li>List top 3 agreed-upon goals.</li> </ul>  |
| Executive summary              | <ul style="list-style-type: none"> <li>Emphasize key messages.</li> </ul>   | <ul style="list-style-type: none"> <li>Emphasize key messages.</li> </ul>  |
| Approach                       | <ul style="list-style-type: none"> <li>Discuss high-level methodology.</li> </ul>   | <ul style="list-style-type: none"> <li>Discuss high-level methodology.</li> <li>Provide relevant details on modeling techniques and technology.</li> </ul>   |
| Model description              | <ul style="list-style-type: none"> <li>Provide overview of the modeling techniques.</li> </ul>                                  | <ul style="list-style-type: none"> <li>Provide overview of the modeling techniques.</li> </ul>   |
| Key points supported with data | <ul style="list-style-type: none"> <li>Support key points with simple charts, and graphics—for example, a bar chart.</li> </ul> | <ul style="list-style-type: none"> <li>To support the key points, show details.</li> <li>Use analyst-oriented charts and graphs, such as ROC curves and histograms.</li> <li>Show visuals of key variables, and discuss significance of each.</li> </ul> |

## Summary of core components for target audiences (cont.)

| Summary of core components for target audiences |  |   |
|---|--|---|
| Presentation component                          | Project sponsor presentation   | Analyst presentation  |
| Model details                                   | <ul style="list-style-type: none"><li>- Omit this section, or discuss only at a high level.</li></ul>  | <ul style="list-style-type: none"><li>- Show more logic or expression of the model.</li><li>- Provide any variables and associated coefficients.</li><li>- Discuss the technologies or tools used to execute the model.</li><li>- Describe expected model performance, and any caveats.</li></ul> |
| Recommendations                                 | <ul style="list-style-type: none"><li>- Focus on business impact of doing this, including risks and ROI.</li><li>- Give the sponsor salient points to help them evangelize the work within the organization.</li></ul> | <ul style="list-style-type: none"><li>- Supplement recommendations with any implications for the modeling, or for deploying in a production environment.</li></ul>  |

## Lesson: Preparing project presentations

| Presentation component | Project sponsor presentation   | Analyst presentation   |
|------------------------|--|--|
| Model details          | <ul style="list-style-type: none"><li>Omit this section, or discuss only at a high level.</li></ul>  | <ul style="list-style-type: none"><li>Show main logic or expression of the model.</li><li>Provide any variables and associated coefficients.</li><li>Discuss the technologies or tools used to execute the model.</li><li>Describe expected model performance and any caveats.</li></ul> |
| Recommendations        | <ul style="list-style-type: none"><li>Focus on business impact of doing this, including risks and ROI.</li><li>Give the sponsor salient points to help them evangelize the work within the organization.</li></ul> | <ul style="list-style-type: none"><li>Supplement recommendations with any implications for the modeling, or for deploying in a production environment.</li></ul>   |

## Final considerations for preparing final presentation

### Final considerations for preparing final presentation

- Be visual, and tell them what you want to show them.
- Be Mutually Exclusive and Collectively Exhaustive (MECE).
- Tie your ideas together; do not force people to tie your ideas together.
- Guide people, and help them draw logical connections.
- Do not forget that not everyone has gone through the discovery phase as you have.
- Context is key. Orient people to the project itself, the graphics you use, the terminology, and the jargon, and spell out acronyms.

101

DATA SCIENCE

DELL EMC

Here is a list of tips and tricks to use, as well as some common pitfalls to avoid, when creating presentations. One of the biggest mistakes authors make when creating presentations is forgetting to take time to set the context in presentations and in presenting the findings. When doing this, keep in mind that you always need to orient the viewer to the work you have created. This is true when writing your goals and situational overview for the project as a whole, and also for creating charts to give the right amount of context to orient the viewer to your main message. Context is key in conveying information in a way the audience can easily understand.

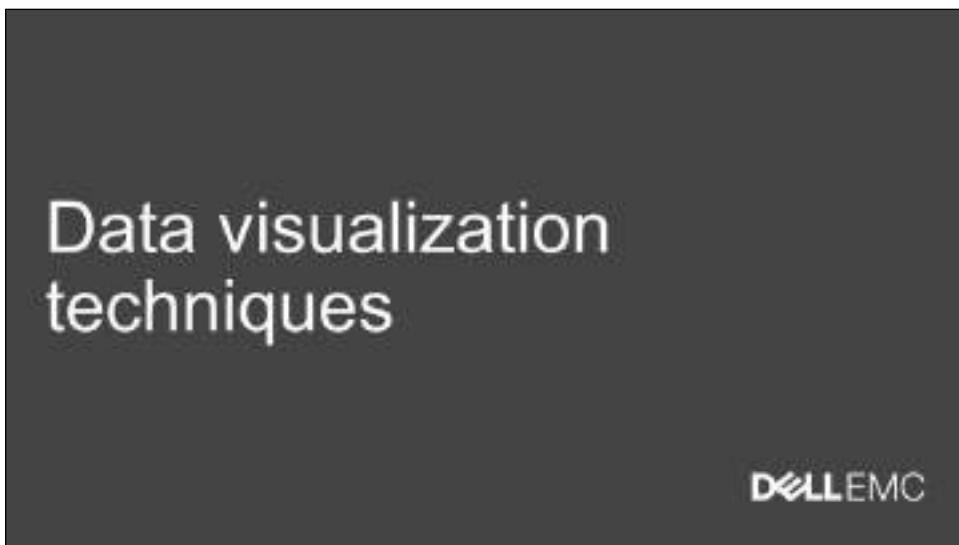
## Final considerations for preparing final presentation (cont.)

### Final considerations for preparing final presentation (cont.)

- Do not assume that people see the obvious benefits.
- Measure and quantify the benefits. Be specific. "\$8.5M in annual cost savings" is much stronger than "Great value."
- Be patient; you may be asked to tell your story more than once. Consider these sessions opportunities to refine your message and share good work that was completed.
- Let the intended audience guide you in shaping the right message and level of detail.
- Avoid long bulleted lists.

## Lesson: Data visualization techniques

### Introduction



### Data visualization techniques

This lesson covers:

- The importance of data visualization
- The iterative nature of building visualizations
- Common visual representation methods
- Cleaning up a graph
- Considerations about using 3D charts

CH 01 INTRODUCTION TO DATA SCIENCE

DELL EMC

This lesson covers the importance of data visualization, the iterative nature of analytics, and common visual representation methods. This lesson also covers cleaning up a graph and considerations about using 3D charts.

# What is data visualization?

|   | What is data visualization?   |
|---|---|
|  | The presentation of statistics with images that depict the meaning of the statistics. – census.gov  |
|  | Data visualization schematically abstracts information to bring about a deeper understanding of the data, wrapping it in an element of awe. – Bloomberg Business Week |
|  | The use of computer-supported, interactive, visual representations of abstract data to amplify cognition. – Card et al., 1999   |

Here are three different definitions of data visualization:



The US Census Bureau defines data visualization as a term that means “the presentation of statistics with images that depict the meaning of the statistics.”<sup>1</sup>

Bloomberg Business Week says, “Data visualization schematically abstracts information to bring about a deeper understanding of the data, wrapping it in an element of awe.”<sup>2</sup>

A better definition is perhaps from Card et al.'s book in 1999. They define data visualization as "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition."<sup>3</sup>

**References:**

<sup>1</sup>Yau, N. Visualizing Census Data. *The United States Census Bureau*.

[https://census.gov/library/video/data\\_visualization1.html](https://census.gov/library/video/data_visualization1.html)

<sup>2</sup>Popova, M. (2012). Data Visualization: Stories for the Information Age.

<https://www.bloomberg.com/news/articles/2009-08-12/data-visualization-stories-for-the-information-age>

<sup>3</sup>Card, S. K., Mackinlay, J. D., & Shneiderman, B. (Eds.). (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

## Data visualization's importance

### Data visualization's importance

**Data**

- Is often easier to process and understand.
- Helps discover new insights.
- Is useful for large datasets with many variables.
- Enables effective communications.



DATA VIZ

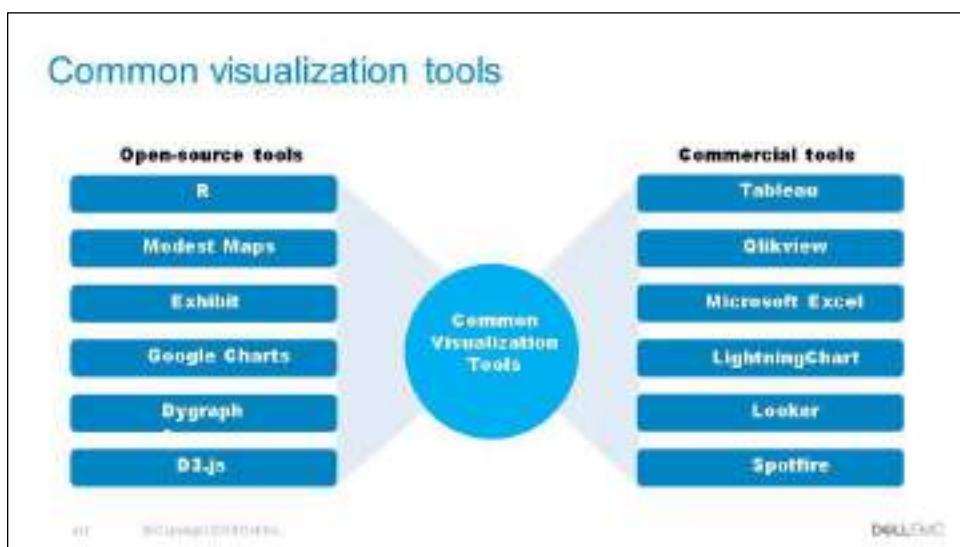
As the volume and complexity of data has grown, users are becoming more reliant on using crisp visuals to illustrate key ideas and also to portray rich data in a digestible way. Although computers have simplified the process to rapidly generate graphs and charts, data visualizations have existed for centuries.

Data is often much easier to process in a visual form that can also provide new insights. Data visualization can be applied to any domain. For example, the same visualization tool – such as a scatterplot – can visualize two variables from any domain, whether we are plotting car fuel efficiency versus the weight of the car, or the country GDP versus the unemployment rate. As the size of the dataset grows and the number of variables increases, the need for effective visualizations increases.

## Common visualization tools

### Discussion Topic: Common visualization tools

Many great visualization tools are on the market to help in creating clear graphics for presentations and applications. This is a listing of some open-source and commercial tools. Some tools are standalone products, but other visualization tools are incorporated into analytic packages, such as SAS or SPSS. The choice of visualization tool should be based on what is to be accomplished. For example, a strictly GUI-driven tool may be useful for prototyping charts and data exploration, but may not be useful when automation is required.



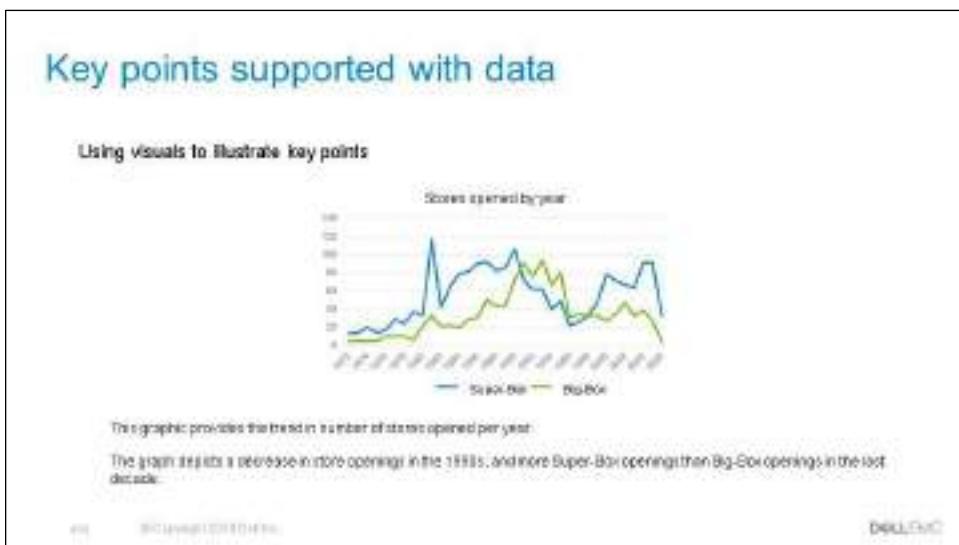
## Communicating key points supported with data

Communicating key points supported with data

35 years of the Big-Box and Super-Box new store data

| Year  | Big-Box | Super-Box | Total |
|-------|---------|-----------|-------|
| 1980  | 11      | 10        | 21    |
| 1981  | 12      | 11        | 23    |
| 1982  | 13      | 12        | 25    |
| 1983  | 14      | 13        | 27    |
| 1984  | 15      | 14        | 29    |
| 1985  | 16      | 15        | 31    |
| 1986  | 17      | 16        | 33    |
| 1987  | 18      | 17        | 35    |
| 1988  | 19      | 18        | 37    |
| 1989  | 20      | 19        | 39    |
| 1990  | 21      | 20        | 41    |
| 1991  | 22      | 21        | 43    |
| 1992  | 23      | 22        | 45    |
| 1993  | 24      | 23        | 47    |
| 1994  | 25      | 24        | 49    |
| 1995  | 26      | 25        | 51    |
| 1996  | 27      | 26        | 53    |
| 1997  | 28      | 27        | 55    |
| 1998  | 29      | 28        | 57    |
| 1999  | 30      | 29        | 59    |
| 2000  | 31      | 30        | 61    |
| 2001  | 32      | 31        | 63    |
| 2002  | 33      | 32        | 65    |
| 2003  | 34      | 33        | 67    |
| 2004  | 35      | 34        | 69    |
| 2005  | 36      | 35        | 71    |
| 2006  | 37      | 36        | 73    |
| 2007  | 38      | 37        | 75    |
| 2008  | 39      | 38        | 77    |
| 2009  | 40      | 39        | 79    |
| 2010  | 41      | 40        | 81    |
| 2011  | 42      | 41        | 83    |
| 2012  | 43      | 42        | 85    |
| 2013  | 44      | 43        | 87    |
| 2014  | 45      | 44        | 89    |
| 2015  | 46      | 45        | 91    |
| 2016  | 47      | 46        | 93    |
| 2017  | 48      | 47        | 95    |
| 2018  | 49      | 48        | 97    |
| 2019  | 50      | 49        | 99    |
| 2020  | 51      | 50        | 101   |
| 2021  | 52      | 51        | 103   |
| 2022  | 53      | 52        | 105   |
| 2023  | 54      | 53        | 107   |
| 2024  | 55      | 54        | 109   |
| 2025  | 56      | 55        | 111   |
| 2026  | 57      | 56        | 113   |
| 2027  | 58      | 57        | 115   |
| 2028  | 59      | 58        | 117   |
| 2029  | 60      | 59        | 119   |
| 2030  | 61      | 60        | 121   |
| 2031  | 62      | 61        | 123   |
| 2032  | 63      | 62        | 125   |
| 2033  | 64      | 63        | 127   |
| 2034  | 65      | 64        | 129   |
| 2035  | 66      | 65        | 131   |
| 2036  | 67      | 66        | 133   |
| 2037  | 68      | 67        | 135   |
| 2038  | 69      | 68        | 137   |
| 2039  | 70      | 69        | 139   |
| 2040  | 71      | 70        | 141   |
| 2041  | 72      | 71        | 143   |
| 2042  | 73      | 72        | 145   |
| 2043  | 74      | 73        | 147   |
| 2044  | 75      | 74        | 149   |
| 2045  | 76      | 75        | 151   |
| 2046  | 77      | 76        | 153   |
| 2047  | 78      | 77        | 155   |
| 2048  | 79      | 78        | 157   |
| 2049  | 80      | 79        | 159   |
| 2050  | 81      | 80        | 161   |
| 2051  | 82      | 81        | 163   |
| 2052  | 83      | 82        | 165   |
| 2053  | 84      | 83        | 167   |
| 2054  | 85      | 84        | 169   |
| 2055  | 86      | 85        | 171   |
| 2056  | 87      | 86        | 173   |
| 2057  | 88      | 87        | 175   |
| 2058  | 89      | 88        | 177   |
| 2059  | 90      | 89        | 179   |
| 2060  | 91      | 90        | 181   |
| 2061  | 92      | 91        | 183   |
| 2062  | 93      | 92        | 185   |
| 2063  | 94      | 93        | 187   |
| 2064  | 95      | 94        | 189   |
| 2065  | 96      | 95        | 191   |
| 2066  | 97      | 96        | 193   |
| 2067  | 98      | 97        | 195   |
| 2068  | 99      | 98        | 197   |
| 2069  | 100     | 99        | 199   |
| 2070  | 101     | 100       | 201   |
| 2071  | 102     | 101       | 203   |
| 2072  | 103     | 102       | 205   |
| 2073  | 104     | 103       | 207   |
| 2074  | 105     | 104       | 209   |
| 2075  | 106     | 105       | 211   |
| 2076  | 107     | 106       | 213   |
| 2077  | 108     | 107       | 215   |
| 2078  | 109     | 108       | 217   |
| 2079  | 110     | 109       | 219   |
| 2080  | 111     | 110       | 221   |
| 2081  | 112     | 111       | 223   |
| 2082  | 113     | 112       | 225   |
| 2083  | 114     | 113       | 227   |
| 2084  | 115     | 114       | 229   |
| 2085  | 116     | 115       | 231   |
| 2086  | 117     | 116       | 233   |
| 2087  | 118     | 117       | 235   |
| 2088  | 119     | 118       | 237   |
| 2089  | 120     | 119       | 239   |
| 2090  | 121     | 120       | 241   |
| 2091  | 122     | 121       | 243   |
| 2092  | 123     | 122       | 245   |
| 2093  | 124     | 123       | 247   |
| 2094  | 125     | 124       | 249   |
| 2095  | 126     | 125       | 251   |
| 2096  | 127     | 126       | 253   |
| 2097  | 128     | 127       | 255   |
| 2098  | 129     | 128       | 257   |
| 2099  | 130     | 129       | 259   |
| 20100 | 131     | 130       | 261   |
| 20101 | 132     | 131       | 263   |
| 20102 | 133     | 132       | 265   |
| 20103 | 134     | 133       | 267   |
| 20104 | 135     | 134       | 269   |
| 20105 | 136     | 135       | 271   |
| 20106 | 137     | 136       | 273   |
| 20107 | 138     | 137       | 275   |
| 20108 | 139     | 138       | 277   |
| 20109 | 140     | 139       | 279   |
| 20110 | 141     | 140       | 281   |
| 20111 | 142     | 141       | 283   |
| 20112 | 143     | 142       | 285   |
| 20113 | 144     | 143       | 287   |
| 20114 | 145     | 144       | 289   |
| 20115 | 146     | 145       | 291   |
| 20116 | 147     | 146       | 293   |
| 20117 | 148     | 147       | 295   |
| 20118 | 149     | 148       | 297   |
| 20119 | 150     | 149       | 299   |
| 20120 | 151     | 150       | 301   |
| 20121 | 152     | 151       | 303   |
| 20122 | 153     | 152       | 305   |
| 20123 | 154     | 153       | 307   |
| 20124 | 155     | 154       | 309   |
| 20125 | 156     | 155       | 311   |
| 20126 | 157     | 156       | 313   |
| 20127 | 158     | 157       | 315   |
| 20128 | 159     | 158       | 317   |
| 20129 | 160     | 159       | 319   |
| 20130 | 161     | 160       | 321   |
| 20131 | 162     | 161       | 323   |
| 20132 | 163     | 162       | 325   |
| 20133 | 164     | 163       | 327   |
| 20134 | 165     | 164       | 329   |
| 20135 | 166     | 165       | 331   |
| 20136 | 167     | 166       | 333   |
| 20137 | 168     | 167       | 335   |
| 20138 | 169     | 168       | 337   |
| 20139 | 170     | 169       | 339   |
| 20140 | 171     | 170       | 341   |
| 20141 | 172     | 171       | 343   |
| 20142 | 173     | 172       | 345   |
| 20143 | 174     | 173       | 347   |
| 20144 | 175     | 174       | 349   |
| 20145 | 176     | 175       | 351   |
| 20146 | 177     | 176       | 353   |
| 20147 | 178     | 177       | 355   |
| 20148 | 179     | 178       | 357   |
| 20149 | 180     | 179       | 359   |
| 20150 | 181     | 180       | 361   |
| 20151 | 182     | 181       | 363   |
| 20152 | 183     | 182       | 365   |
| 20153 | 184     | 183       | 367   |
| 20154 | 185     | 184       | 369   |
| 20155 | 186     | 185       | 371   |
| 20156 | 187     | 186       | 373   |
| 20157 | 188     | 187       | 375   |
| 20158 | 189     | 188       | 377   |
| 20159 | 190     | 189       | 379   |
| 20160 | 191     | 190       | 381   |
| 20161 | 192     | 191       | 383   |
| 20162 | 193     | 192       | 385   |
| 20163 | 194     | 193       | 387   |
| 20164 | 195     | 194       | 389   |
| 20165 | 196     | 195       | 391   |
| 20166 | 197     | 196       | 393   |
| 20167 | 198     | 197       | 395   |
| 20168 | 199     | 198       | 397   |
| 20169 | 200     | 199       | 399   |
| 20170 | 201     | 200       | 401   |
| 20171 | 202     | 201       | 403   |
| 20172 | 203     | 202       | 405   |
| 20173 | 204     | 203       | 407   |
| 20174 | 205     | 204       | 409   |
| 20175 | 206     | 205       | 411   |
| 20176 | 207     | 206       | 413   |
| 20177 | 208     | 207       | 415   |
| 20178 | 209     | 208       | 417   |
| 20179 | 210     | 209       | 419   |
| 20180 | 211     | 210       | 421   |
| 20181 | 212     | 211       | 423   |
| 20182 | 213     | 212       | 425   |
| 20183 | 214     | 213       | 427   |
| 20184 | 215     | 214       | 429   |
| 20185 | 216     | 215       | 431   |
| 20186 | 217     | 216       | 433   |
| 20187 | 218     | 217       | 435   |
| 20188 | 219     | 218       | 437   |
| 20189 | 220     | 219       | 439   |
| 20190 | 221     | 220       | 441   |
| 20191 | 222     | 221       | 443   |
| 20192 | 223     | 222       | 445   |
| 20193 | 224     | 223       | 447   |
| 20194 | 225     | 224       | 449   |
| 20195 | 226     | 225       | 451   |
| 20196 | 227     | 226       | 453   |
| 20197 | 228     | 227       | 455   |
| 20198 | 229     | 228       | 457   |
| 20199 | 230     | 229       | 459   |
| 20200 | 231     | 230       | 461   |
| 20201 | 232     | 231       | 463   |
| 20202 | 233     | 232       | 465   |
| 20203 | 234     | 233       | 467   |
| 20204 | 235     | 234       | 469   |
| 20205 | 236     | 235       | 471   |
| 20206 | 237     | 236       | 473   |
| 20207 | 238     | 237       | 475   |
| 20208 | 239     | 238       | 477   |
| 20209 | 240     | 239       | 479   |
| 20210 | 241     | 240       | 481   |
| 20211 | 242     | 241       | 483   |
| 20212 | 243     | 242       | 485   |
| 20213 | 244     | 243       | 487   |
| 20214 | 245     | 244       | 489   |
| 20215 | 246     | 245       | 491   |
| 20216 | 247     | 246       | 493   |
| 20217 | 248     | 247       | 495   |
| 20218 | 249     | 248       | 497   |
| 20219 | 250     | 249       | 499   |
| 20220 | 251     | 250       | 501   |
| 20221 | 252     | 251       | 503   |
| 20222 | 253     | 252       | 505   |
| 20223 | 254     | 253       | 507   |
| 20224 | 255     | 254       | 509   |
| 20225 | 256     | 255       | 511   |
| 20226 | 257     | 256       | 513   |
| 20227 | 258     | 257       | 515   |
| 20228 | 259     | 258       | 517   |
| 20229 | 260     | 259       | 519   |
| 20230 | 261     | 260       | 521   |
| 20231 | 262     | 261       | 523   |
| 20232 | 263     | 262       | 525   |
| 20233 | 264     | 263       | 527   |
| 20234 | 265     | 264       | 529   |
| 20235 | 266     | 265       | 531   |
| 20236 | 267     | 266       | 533   |
| 20237 | 268     | 267       | 535   |
| 20238 | 269     | 268       | 537   |
| 20239 | 270     | 269       | 539   |
| 20240 | 271     | 270       | 541   |
| 20241 | 272     | 271       | 543   |
| 20242 | 273     | 272       | 545   |
| 20243 | 274     | 273       | 547   |
| 20244 | 275     | 274       | 549   |
| 20245 | 276     | 275       | 551   |
| 20246 | 277     | 276       | 553   |
| 20247 | 278     | 277       | 555   |
| 20248 | 279     | 278       | 557   |
| 20249 | 280     | 279       | 559   |
| 20250 | 281     | 280       | 561   |
| 20251 | 282     | 281       | 563   |
| 20252 | 283     | 282       | 565   |
| 20253 | 284     | 283       | 567   |
| 20254 | 285     | 284       | 569   |
| 20255 | 286     | 285       | 571   |
| 20256 | 287     | 286       | 573   |
| 20257 | 288     | 287       | 575   |
| 20258 | 289     | 288       | 577   |
| 20259 | 290     | 289       | 579   |
| 20260 | 291     | 290       | 581   |
| 20261 | 292     | 291       | 583   |
| 20262 | 293     | 292       | 585   |
| 20263 | 294     | 293       | 587   |
| 20264 | 295     | 294       | 589   |
| 20265 | 296     | 295       | 591   |
| 20266 | 297     | 296       | 593   |
| 20267 | 298     | 297       | 595   |
| 20268 | 299     | 298       | 597   |
| 20269 | 300     | 299       | 599   |
| 20270 | 301     | 300       | 601   |
| 20271 | 302     | 301       | 603   |
| 20272 | 303     | 302       | 605   |
| 20273 | 304     | 303       | 607   |
| 20274 | 305     | 304       | 609   |
| 20275 | 306     | 305       | 611   |
| 20276 | 307     | 306       | 613   |
| 20277 | 308     | 307       | 615   |
| 20278 | 309     | 308       | 617   |
| 20279 | 310     | 309       | 619   |
| 20280 | 311     | 310       | 621   |
| 20281 | 312     | 311       | 623   |
| 20282 | 313     | 312       | 625   |
| 20283 | 314     | 313       | 627   |

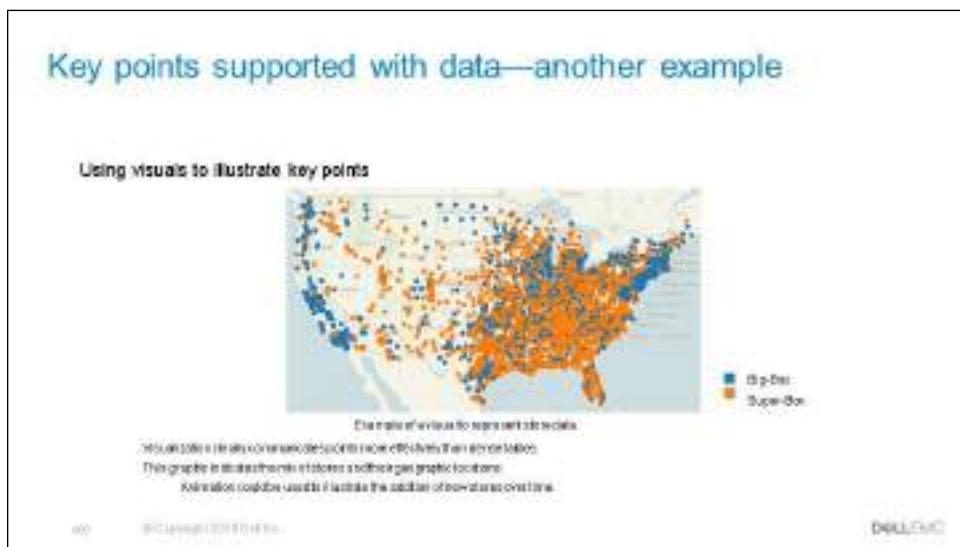
## Key points supported with data



The figure on this slide shows the previous table graphically. The graph clearly indicates the variation in the number of stores opened, and it is easier to interpret than the table.

Here, we can easily see that there is a significant drop in the last few years in number of stores opened.

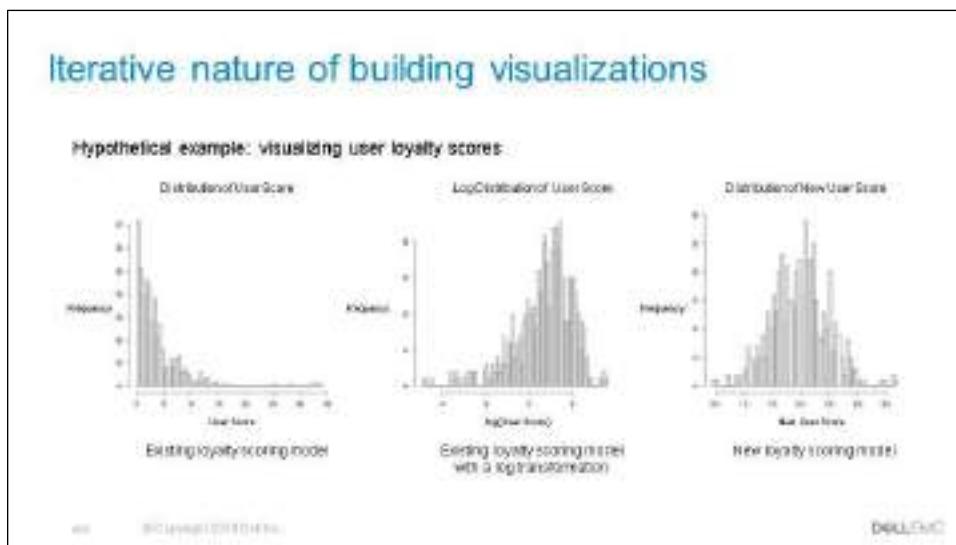
## Key points supported with data—another example



This image is a map of the United States, with the points representing the geographic locations of the stores. This map is a more powerful way to depict data than a small table would be. The approach is well suited to a sponsor audience.

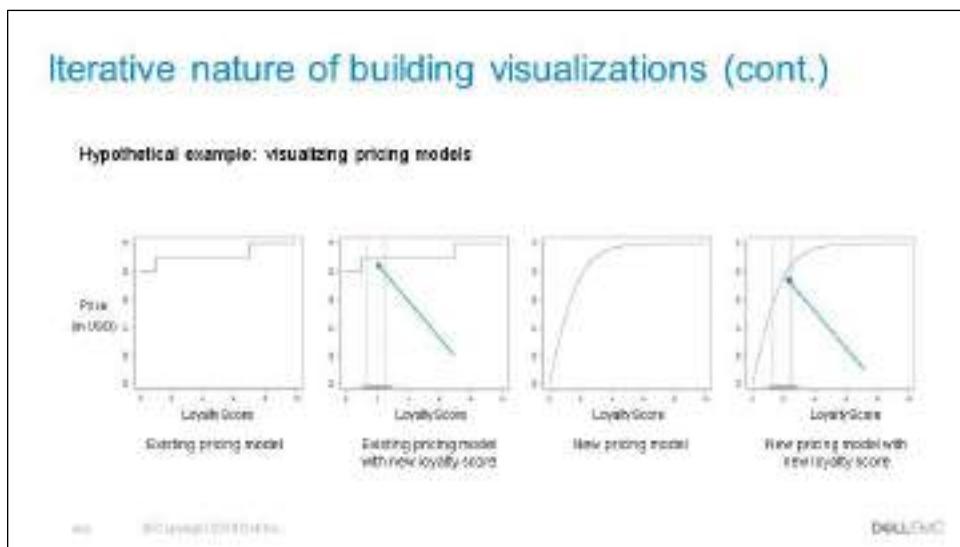
The map shows where the company has market saturation and where the company has grown. During a presentation, it is still important to clearly articulate verbally or on the slide the key message of the visualization: for example, “Big-Box stores are concentrated on the Northeast and the West coasts.”

## Iterative nature of building visualizations



These graphs portray a hypothetical example of some of the steps a data scientist may go through in analyzing user loyalty scores. Data scientists typically iterate and view the data in many different ways, framing hypotheses, testing them, and exploring the implications of a given model. In this case, we examine the distribution of loyalty scores for an existing model as well as a new model.

## Iterative nature of building visualizations (cont.)



Building on the previous example, the new loyalty scoring model results are overlaid on the existing and proposed pricing models. As shown in the second graph, almost all customers receive the same price tiering, regardless of their loyalty score. With the new pricing model, the least loyal customers receive a bigger incentive.

## Presenting pricing model results to sponsor

### Presenting pricing model results to sponsor

- Before the project, pricing promotions were offered to all customers equally.
- With the new approach for pricing promotions:
  - Highly loyal customers do not receive as many price promotions.
  - Customers with low loyalty receive more price promotions to influence their buying decisions.
- State the benefits of the new approach:
  - Avoid \$2M in lost customer business
  - \$1.5M reduction in new customer acquisition expenditures
  - \$1M savings in unnecessary pricing promotions

40

DATA SCIENCE

Dell EMC

Here is an example of the output from the price optimization project scenario, showing how one may present this to an audience of project sponsors. This shows a simple bar chart to depict the average price per customer or user segment. This is a much simpler-looking visual than the prior slide, and this one clearly shows that customers with lower loyalty scores tend to get lower prices, due to targeting from price promotions.

Note that the comments at the left of the graphic relate to explaining the impact of the model at a high level and the cost savings of implementing this approach to price optimization.

## Choosing correct chart type

| Types of information | Recommended charts                   |
|----------------------|--------------------------------------|
| Components           | Pie chart                            |
| Bar                  | Bar chart                            |
| Time series          | Line chart                           |
| Frequency            | Line charts, histograms              |
| Correlation          | Scatterplot, side-by-side bar charts |

Some recommended chart types are provided based on the type of data to be presented. This table is by no means exhaustive. Consider the message you are trying to communicate, and then choose an appropriate visual to support the point. Misuse of charts tends to confuse the audience, so be sure to take into account the data type and message when choosing a chart.

**Pie charts** are designed to show the components, or parts, relative to the whole set of things. This is also the most widely used chart. If you are going to use a pie chart, use it when showing only two to three items in a chart and only for sponsor audiences.

**Bar charts and line charts** are used more often, and are very useful for showing comparisons and trends over time. For bar charts, horizontal bar charts allow you to fit the text labels better and provide more horizontal space to fit them next to a chart, even though many people tend to use vertical bar charts. Vertical bar charts tend to work well when the labels are small, such as when showing comparisons over time, using years.

| Types of information | Recommended charts                   |
|----------------------|--------------------------------------|
| Components           | Pie chart                            |
| Item                 | Bar chart                            |
| Time series          | Line chart                           |
| Frequency            | Line charts, histograms              |
| Correlation          | Scatterplot, side-by-side bar charts |

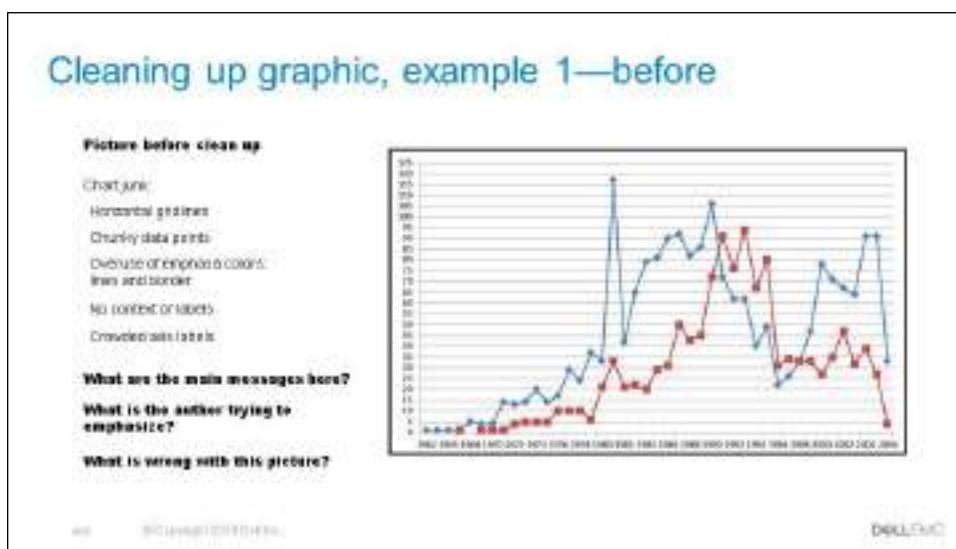
For frequency, **histograms** will show the distribution of data, and are useful for showing information to an analyst audience. The data distributions are typically one of the first steps in visualization data to prepare for the model planning. When examining possible correlations, scatterplots are useful to compare relationships among variables.

As with any presentation, consider the audience and their level of expertise when selecting the chart to convey your message. These charts are simple examples, but can easily become more complex with additional data variables, combining charts together, or adding animation where appropriate. Regardless of the chart type selected, it is important to select the proper choice of colors, shading, markers, and so on to effectively bring the audience's attention to the relevant data.

## Cleaning up graphic, example 1—before

### Discussion Topic: Cleaning up graphic, example 1—before

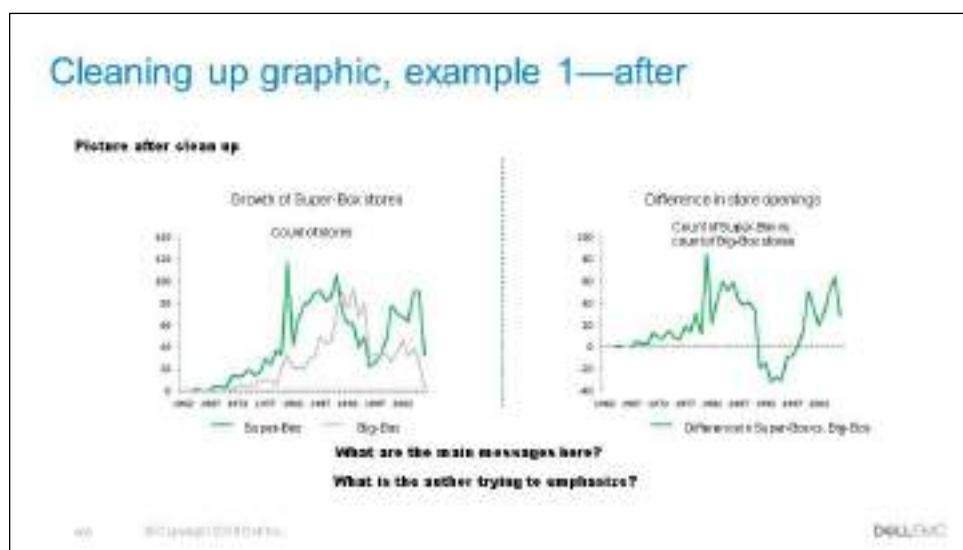
The image on this slide is an example of a line chart comparing two trends over time. It's a busy-looking chart and contains a lot of "chart junk," which distracts the viewer from the main message. The points mentioned on the slide are some of the chart junk this visual suffers from, which is easily addressed as shown in the next slide. Note that there is no clear message associated with this chart and no legend to provide context for what is shown.



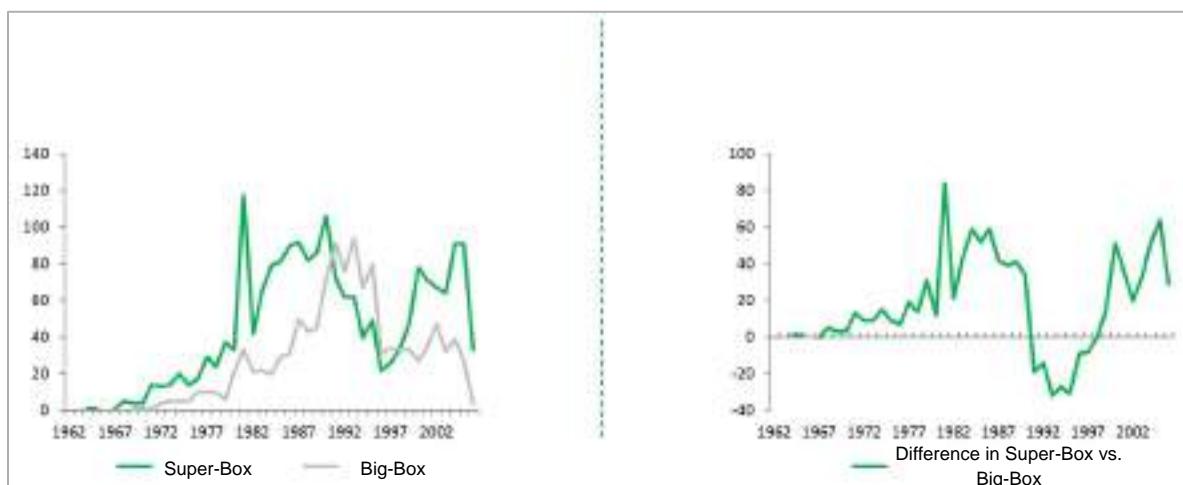
## Cleaning up graphic, example 1—after

### Discussion Topic: Cleaning up graphic, example 1—after

These are two examples of cleaned-up versions of the chart on the previous page. Note that the problems with chart junk have been addressed, there is a **clear label and title for each chart to reinforce the message, and color has been used in ways to highlight the point the author is trying to make.**



Note the amount of white space being used in each of the two charts. Removing gridlines, excessive axes, and the visual noise within the chart allows you to create very clear contrast between emphasis colors and the standard colors—in this example, the green lines as opposed to the light gray.



## Lesson: Data visualization techniques

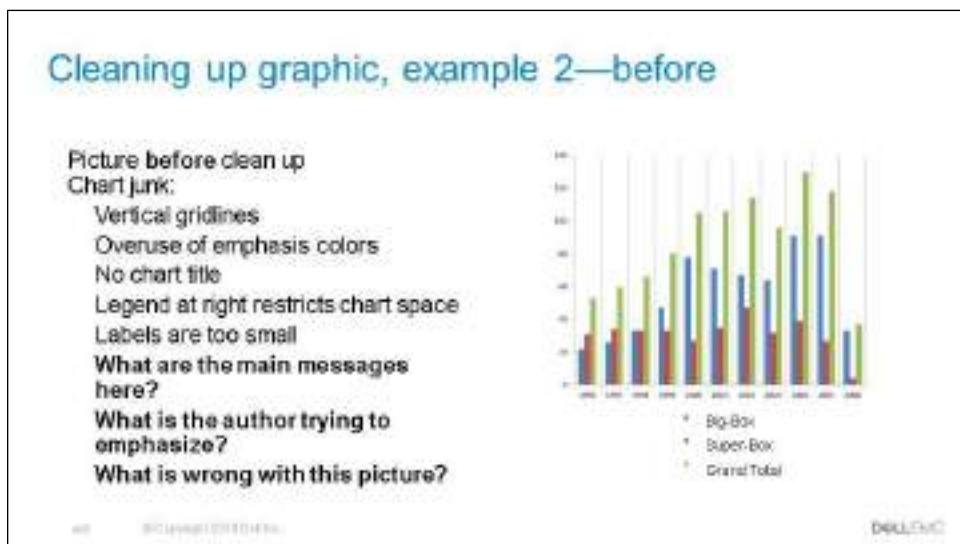
When creating charts, it is best to do most of your main visuals in standard colors, light tones, or color shades so that you can choose to add stronger emphasis colors to emphasize the main points and draw attention to the parts of the graphic that demonstrate your main points. In this case, we have made the trend of Big-Box stores in light gray to fade into the background, but not disappear, while making the Super-Box stores trend in a bright green and making it prominent to support the point the author is making about the growth of the Super-Box stores.

An alternative is shown at right. If the main message is to show the difference in the growth of new stores, you could simplify the chart further and choose to graph only the difference between Super-Box stores compared to regular Big-Box stores. **Two examples are shown to illustrate different ways to convey your message, depending on what it is you would like to show.**

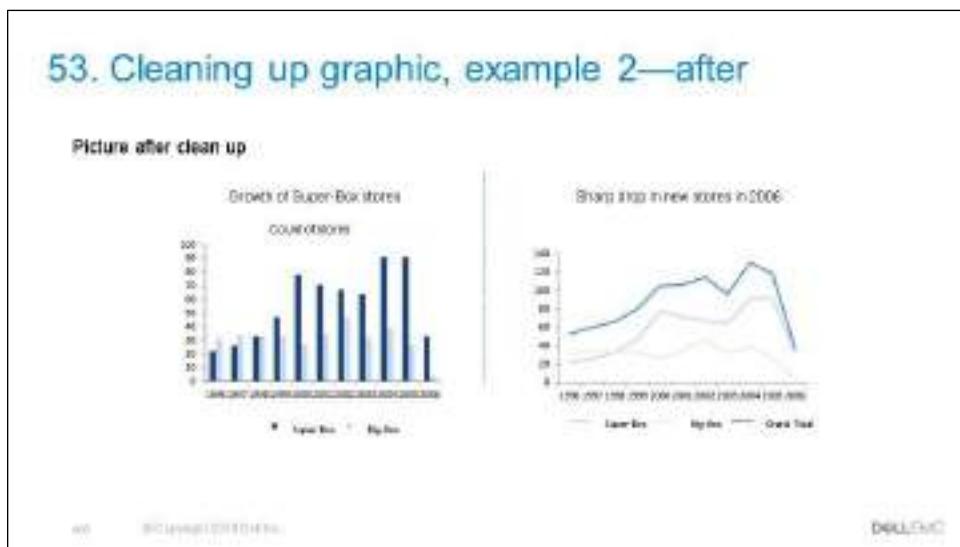
## Cleaning up graphic, example 2—before

### Discussion Topic: Cleaning up graphic, example 2—before

Here is a sample graph with the typical problems related to chart junk, including misuse of color schemes, and lack of context. Shown at left are the main problems with the **graph**, with cleaned-up alternatives to this visual on the subsequent pages.



## Cleaning up graphic, example 2—after

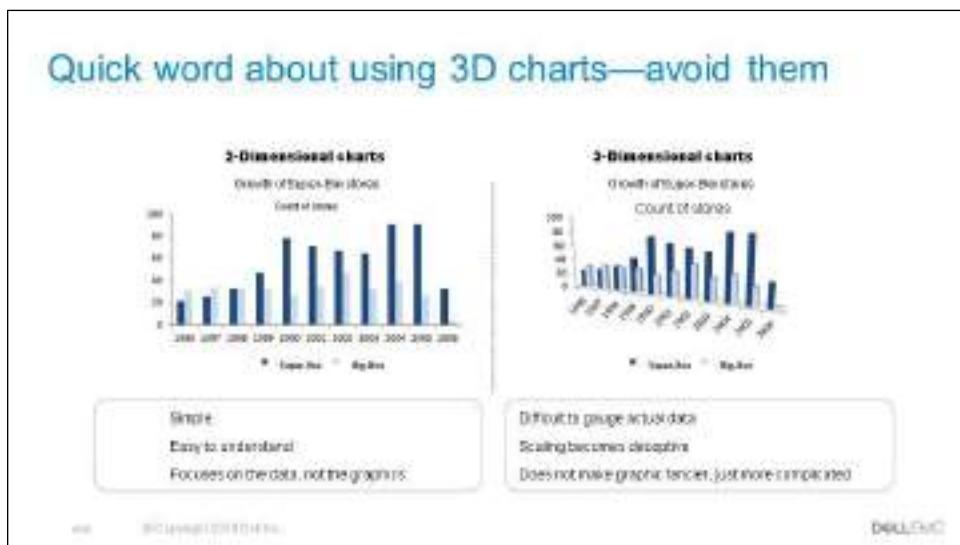


The graphs shown on the slide are some simplified and cleaned-up versions of the previous slide's graphic. These graphs show two options for modifying the graphic, depending on the main point the presenter is trying to make.

The chart on the left of the slide shows a strong, emphasis color – dark blue – representing the Super-Box stores, to support the chart title about the growth of Super-Box stores. If the presenter wanted instead to talk about the total growth of stores, a line chart showing the trends over time – as shown on the right – would be a better choice.

In both cases, these graphs remove the noise and distractions within the chart, de-emphasize the data that does not contribute to the key point, and make prominent the data that reinforces the key point as stated in the chart's title.

## Quick word about using 3D charts—avoid them



The images shown on this slide are a side-by-side comparison of two charts. As mentioned, 3-dimensional charts often distort scales and axes, and impede viewer cognition. The charts on the left and right portray the same data; however, when looking at 3-dimensional charts, it is more difficult to judge the actual height of the bars. In addition, the shadowing and shape of the chart cause most viewers to spend time looking at the perspective of the chart, rather than the height of the bars, which is the key message and purpose of this visual.

## Tips for building effective data visualizations

### Tips for building effective data visualizations

#### Remove distraction:

- Minimize chart junk
- Data-ink ratio

#### Choose the simplest, clearest visual for the situation:

- Strive to illustrate your points
- Charts should reinforce your key messages
- Charts vs. data art

#### Use color deliberately:

- Emphasis colors vs. standard colors
- Usually, less is more
- Focus on the contrast

#### Context:

- Consistent scales, labels, axes
- Use logs vs. raw values to show differences

44

DATA VISUALIZATION

DELL EMC

The points listed here summarize many of the key ideas in the preceding examples. Follow these suggestions to minimize distractions in the visualizations and to focus on clearly communicating the key messages.

Similar to the idea of removing chart junk is being cognizant of the data-ink ratio. Data-ink refers to the actual portion of a graphic that is used to portray the data itself, while non-data ink represents labels, edges, colors, and other decoration. The ratio = (data-ink)/(total ink used to print the graphic). In other words, the greater the ratio of data-ink in your visual, the more data rich it is and the fewer distractions it has.

Context is critical to orient the viewer of a visualization, as people have immediate reactions to imagery on a pre-cognitive level. To this end, be sure to make thoughtful use of color, and orient the viewer with scales, legends, and axes.

## Check your knowledge

Check your knowledge

Which chart is suitable to represent time series data?

A. Line chart      C. Pie chart  
B. Bar chart      D. Scatterplot

40 © 2018 Dell Inc.

Dell EMC

### Check your knowledge:

Which chart is suitable to represent time series data?

- A. Line chart
- B. Bar chart
- C. Pie chart
- D. Scatterplot

### Question:

Which chart is suitable to represent time series data?

### Answer:

---

