DOWNLOAD PDF MODERN DATA ANALYSIS

Chapter 1 : modern data sources | Analysis Services Team Blog

Modern Data Analysis contains the proceedings of a Workshop on Modern Data Analysis held in Raleigh, North Carolina, on June , under the auspices of the United States Army Research Office.

Retrieve Value Given a set of specific cases, find attributes of those cases. What is the value of aggregation function F over a given set S of data cases? What is the sorted order of a set S of data cases according to their value of attribute A? What is the range of values of attribute A in a set S of data cases? What is the distribution of values of attribute A in a set S of data cases? What is the correlation between attributes X and Y over a given set S of data cases? Barriers to effective analysis[edit] Barriers to effective analysis may exist among the analysts performing the data analysis or among the audience. Distinguishing fact from opinion, cognitive biases, and innumeracy are all challenges to sound data analysis. Confusing fact and opinion[edit] You are entitled to your own opinion, but you are not entitled to your own facts. Daniel Patrick Moynihan Effective analysis requires obtaining relevant facts to answer questions, support a conclusion or formal opinion , or test hypotheses. Facts by definition are irrefutable, meaning that any person involved in the analysis should be able to agree upon them. This makes it a fact. Whether persons agree or disagree with the CBO is their own opinion. As another example, the auditor of a public company must arrive at a formal opinion on whether financial statements of publicly traded corporations are "fairly stated, in all material respects. When making the leap from facts to opinions, there is always the possibility that the opinion is erroneous. Cognitive biases[edit] There are a variety of cognitive biases that can adversely affect analysis. In addition, individuals may discredit information that does not support their views. Analysts may be trained specifically to be aware of these biases and how to overcome them. In his book Psychology of Intelligence Analysis, retired CIA analyst Richards Heuer wrote that analysts should clearly delineate their assumptions and chains of inference and specify the degree and source of the uncertainty involved in the conclusions. He emphasized procedures to help surface and debate alternative points of view. However, audiences may not have such literacy with numbers or numeracy ; they are said to be innumerate. Persons communicating the data may also be attempting to mislead or misinform, deliberately using bad numerical techniques. More important may be the number relative to another number, such as the size of government revenue or spending relative to the size of the economy GDP or the amount of cost relative to revenue in corporate financial statements. This numerical technique is referred to as normalization [7] or common-sizing. There are many such techniques employed by analysts, whether adjusting for inflation i. Analysts apply a variety of techniques to address the various quantitative messages described in the section above. Analysts may also analyze data under different assumptions or scenarios. For example, when analysts perform financial statement analysis , they will often recast the financial statements under different assumptions to help arrive at an estimate of future cash flow, which they then discount to present value based on some interest rate, to determine the valuation of the company or its stock. Smart buildings[edit] A data analytics approach can be used in order to predict energy consumption in buildings. Analytics and business intelligence[edit] Main article: Analytics Analytics is the "extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions. Initial data analysis[edit] The most important distinction between the initial data analysis phase and the main analysis phase, is that during initial data analysis one refrains from any analysis that is aimed at answering the original research question. The initial data analysis phase is guided by the following four questions: Data quality can be assessed in several ways, using different types of analysis: Test for common-method variance. The choice of analyses to assess the data quality during the initial data analysis phase depends on the analyses that will be conducted in the main analysis phase. One should check whether structure of measurement instruments corresponds to structure reported in the literature. There are two ways to assess measurement: If the study did not need or use a randomization procedure, one should check the success of the non-random sampling, for instance by checking whether all subgroups of the population of interest are represented in sample. Other possible data distortions that should be checked are: It is especially important to exactly determine the structure of the sample and specifically the size of the

subgroups when subgroup analyses will be performed during the main analysis phase. The characteristics of the data sample can be assessed by looking at: Basic statistics of important variables Scatter plots Cross-tabulations [31] Final stage of the initial data analysis[edit] During the final stage, the findings of the initial data analysis are documented, and necessary, preferable, and possible corrective actions are taken. Also, the original plan for the main data analyses can and should be specified in more detail or rewritten. In order to do this, several decisions about the main data analyses can and should be made: In the case of non- normals: In the case of missing data: In the case of outliers: In case items do not fit the scale: In the case of too small subgroups: In case the randomization procedure seems to be defective:

Chapter 2 : EECSc - Optimization for Large Scale Data Analysis

This bar-code number lets you verify that you're getting exactly the right version or edition of a book. The digit and digit formats both work.

Nor is the act of planning modern data architectures a technical exercise, subject to the purchase and installation of the latest and greatest shiny new technologies. Rather, the design and creation of modern data architectures is an uplifting process that brings in the whole enterprise, stimulating new ways of thinking, collaborating, and planning for data and information requirements. One thing is clear: Architecture is more important than ever because it provides a road map for the enterprise to follow. Here are the essential components that need to go into building a modern data architecture: To be of value, information needs to have a high business impact. This data may have been within enterprise data environments for some time, but the means and technologies to surface such data, and draw insights, have been prohibitively expensive. The process of identifying, ingesting, and building models for data needs to assure quality and relevance for the business. If a new solution comes on the market â€”the way NoSQL arose a few years backâ€”the architecture should be able to accommodate it. The types of data coming into enterprises can change, as do the tools and platforms that are put into place to handle them. The key is to design a data environment that can accommodate such change. There is the need to facilitate real-time access to data, which could be historical; and there is the requirement to support data from events as they are occurring. For the first category, existing infrastructure such as data warehouses have a critical role to play. For the second, new approaches such as streaming analytics are critical. Data may be coming from transactional applications, as well as devices and sensors across the Internet of Things and mobile devices. These threats are constantly evolvingâ€”they may be coming through email one month, and through flash drives the next. The need for an MDM-based architecture is criticalâ€”organizations are consistently going through changes, including growth, realignments, mergers, and acquisitions. Often, enterprises end up with data systems running in parallel, and often, critical records and information may be duplicated and overlap across these silos. MDM also assures that applications and systems across the enterprise have the same view of a customer, versus disparate or conflicting pieces of data. Access is enabled through a virtualized data services layer that standardizes all data sourcesâ€”regardless of device, applicator, or systems. Data as a service is by definition a form of internal cloud, in that dataâ€”along with accompanying data management platforms, tools, and applicationsâ€”are made available to the enterprise as reusable, standardized services. The potential advantage of data as a service is that processes and assets can be prepackaged based on corporate or compliance standards and made readily available within the enterprise cloud. In the process, data application can reach and serve a larger audience than previous generations of more limited data applications. The route to self-service is providing front-end interfaces that are simply laid out and easy to use for business owners. In the process, a logical service layer can be developed that can be re-used across various projects, departments, and business units. IT still has an important role to play in a self-service-enabled architectureâ€”providing for security, monitoring, and data governance. There is a new generation of tools and templates now available from vendors that enable users to explore datasets with highly visual, even 3D views, that can be adjusted, re-adjusted, and manipulated to look for outliers and trends.

Chapter 3 : "Modern Data Analysis: A First Course in Applied Statistics" by Lawrence C. Hamilton

Modern Data Analysis contains the proceedings of a Workshop on Modern Data Analysis held in Raleigh, North Carolina, on June , under the auspices of the United States Army Research Office. The papers review theories and methods of data analysis and cover topics ranging from single and multiple quantile-quantile (Q-Q) plotting procedures.

Engineering at Betterment Modern Data Analysis: October 29, Companies now are innovating and improving the craft of using data to do business. Companies like Betterment are hiring data scientists and analysts who use software development techniques to reliably answer business questions which have quickly expanded in scale and complexity. To do good data work today, you need to use a system that is reproducible, versionable, scalable, and open. Just as the Ford Motor Company created efficiency with assembly line production and Pixar opened up new worlds by computerizing animation, companies now are innovating and improving the craft of using data to do business. Betterment is one of them. We are built from the ground up on a foundation of data. To avoid time-consuming manual processes, and the human error typical of that approach, analytics has become a programming discipline. With VisiCalc, the first-ever spreadsheet program, in and Excel in , the business world stepped into two new eras in which any employee could manage large amounts of data. The bottlenecks in business analytics had been the speed of human arithmetic or the hours available on corporate mainframes operated by only a few specialists. With spreadsheet software in every cubicle, analytical horsepower was commoditized and Excel jockeys were crowned as the arbiters of truth in business. But the era of the spreadsheet is over. The data is too large, the analyses are too complex, and mistakes are too dangerous to trust to our dear old friend the spreadsheet. Ask Carmen Reinhart and Kenneth Rogoff , two Harvard economists who published an influential paper on sovereign debt and economic growth, only to find out that the results rested in part on the accidental omission of five cells from an average. Requirements for Modern Data Analysis Spreadsheets fundamentally lack these properties essential to modern data work. To do good data work today, you need to use a system that is: That code should take me from the raw data to the conclusions. Most analyses contain too many important detailed steps to plausibly communicate in an email or during a meeting. Reproducible also means efficient. When an input or an assumption changes, it should be as easy as re-running the whole thing. Versionable Code versioning frameworks, such as git, are now a staple in the workflow of most technical teams. Sharing code in a common environment also enables the reuse of modular analysis components. Scalable There are hard technical limits to how large an analysis you can do in a spreadsheet. Excel is capped at just more than 1 million rows. There are also feasibility limits. How long does it take your computer to open a million row spreadsheet? Open Many analyses meet the above ideals but have been produced with expensive, proprietary statistical software that inhibits sharing and reproducibility. If I do an analysis with open-source tools like R or Python, I can post full end-to-end instructions that anyone in the world can reproduce, check, and expand upon. Platforms that introduce compatibility problems between versions and save their data in proprietary formats may limit access to your own work even if you are paying for the privilege. This may seem less important inside a corporate bubble where everyone has access to the same proprietary platform, but it is at the very least a turnoff to most new talent in the field. What to Use, and How Short answer: Here at Betterment, we use both. We use Python more for data pipeline processes and R more for modeling, analyses, and reporting. But this article is not about the relative merits of these popular modern solutions. It is about the merits of using one of them or any of the smaller alternatives. To get the most out of a programmatic data analysis workflow, it should be truly end-to-end, or as close as you can get in your environment. If you are new to one or both of these environments, it can be daunting to sort through all of the tools and figure out what does what. These are some of the most popular tools in each language organized by their layer in your full-stack analysis workflow:

Chapter 4 : Modern methods of data analysis - John Fox, J. Scott Long - Google Books

Trifacta Wrangler is the best modern data analysis tool for preparing and transforming data for analysis. Learn more about Trifacta. NEWS Trifacta Named No. 1 Data Preparation Technology in Ovum Decision Matrix.

Leading enterprises, agile startups, and planet-scale internet companies have all adopted BigQuery with equal ease, and customers from industries as diverse as retail, financial services, healthcare, and gaming are using it to uncover valuable insights from their dataâ€”all at an impressive price-performance to price ratio. With analyst firm Forrester recently recognizing Google as a Leader in their Cloud Data Warehouse industry report, I sat down with Engineering Director Jordan Tigani to talk about the evolution of data warehousing, the current technology landscape, and how BigQuery fits in. What are some of the key engineering choices you and your team made to make BigQuery easy for our customers to adopt? We built BigQuery to serve as a cloud-native data warehouse. Storage and compute are decoupled and can scale independently, on-demand. This is very different from traditional node-based cloud data warehouse solutions or on-premise massively parallel processing MPP systems. Serverless is a simple but powerful concept when it comes to gigabyte- to petabyte-scale data analysis. There is no need for customers to define nodes or clusters. BigQuery also automatically manages query performance based on the volume of data it needs to process. This is a fundamentally different approach. In the recent Forrester Wave for Cloud Data Warehouse, Google received a 5 out of 5 in the performance and scale criterion, the highest score possible among all vendors. In data warehousing, storage is often as important as the analysis engine. I come from a storage background, so I love hearing this question. Clustering allows you to quickly find a needle in a haystack, and you essentially pay the price of the needle, not the haystack. A lot of the enterprises adopting BigQuery are migrating their data warehouse from traditional on-premise systems. Does this create a new set of requirements for your team and how is your team addressing these differences? The ability to handle migrations from on-prem systems is critical for enterprise adoption. And BigQuery supports native integration with enterprise business intelligence BI tools such as Tableau and Looker, and ETL tools such as Informatica, Talend, and Stitch to reduce change management to a great extent for enterprise customers. As I talk to more enterprises, I increasingly hear that data security and reliability are top of mind. How are we addressing these critical needs? BigQuery eliminates the data operations burden by providing automatic data replication for disaster recovery and high availability of processing for no additional charge. BigQuery offers a BigQuery also offers fine-grained identity and access-management controls to make it easier to maintain strong security. Plus, BigQuery data is always encrypted, both at rest and in transit. Traditional data warehouses were designed for the batch analytics paradigm. BigQuery supports both batch and streaming data inserts. What are our plans to support real-time analytics with BigQuery? We find that this trend applies primarily to financial services, e-commerce, gaming and media customers. For example, Zulily uses BigQuery to stream billions of events from their web applications and perform real-time analytics. In the enterprise, data warehouses have primarily been used for BI and reporting applications, but enterprises are increasingly undertaking AI initiatives. What are our plans with BigQuery to support the current and future machine learning-specific needs of our enterprise customers? Our partner engineering team has been working with leading BI partners to offer native integration with BigQuery. But enterprises often expect to retrieve data from their traditional data warehouses for AI and ML projects, and in doing so, they create data silos. This silo effect was a challenge we addressed inside Google, and we knew it was a problem we had to solve for our customers as well. We have evolved BigQuery into a flexible, powerful foundation for machine learning and artificial intelligence. What inspired you and your team to build this functionality, and can you share some details on how BigQuery GIS works? Geo-spatial data is becoming a fundamental component of many customer applications. Especially for customers operating in retail, logistics, and energy sectors. Our customers are already storing geolocation data inside BigQuery tables. We wanted to make analysis on the appropriate data types easy. BigQuery engineering team worked with the Google Earth Engine engineering team to introduce geo-spatial analytics capability directly inside BigQuery. With support for arbitrary points, lines, polygons, and multi-polygons in WKT and GeoJSON

format, data analysts can simplify geospatial analyses, visualize location-based data in new ways, or unlock entirely new lines of business with the power of BigQuery. If we take a step back to think about data warehousing as a product category, what are some of the changes that you are anticipating in the coming few years? Users will spend less time worrying about the shape or size of their data, and more time worrying about what questions they want to ask of their data. Deleting data in order to make it fit into the data warehouse will be a distant memory. I think that streaming sources will be critically important, and that will push people to think about their data in a different way. Cloud Data Warehouse, Q4 report here.

Chapter 5 : Modern data analysis in Excel - Microsoft Tech Community -

Modern data analysis in Excel There is a video from Microsoft website on the above. May i know where to download the file "Product blog.quintoapp.com" and "Sales by blog.quintoapp.com"?

Chapter 6 : Modern data analysis landing page concept Vector | Free Download

Modern Methods of Data Analysis - WS 07/08 Stephanie Hansmann-Menzemer The Cheating Baker Once upon a time, in a holiday resort the landlord L. ran a.

Chapter 7 : Data Best Practices: Don't Trust Your Spreadsheet - Betterment

Just looking at your list above, to fully take advantage of the data, you will want someone who understands data mining, another with predictive analytics background, possibly an R expert, and someone with a high level of analysis capability.

Chapter 8 : Data analysis - Wikipedia

Modern data systems still mainly process data in batch. The next stage is to move to "real time" technologies and make the entire company operate on an "event" instead of the year, the.

Chapter 9 : 8 Steps to Building a Modern Data Architecture - Database Trends and Applications

Companies now are innovating and improving the craft of using data to do business. Companies like Betterment are hiring data scientists and analysts who use software development techniques to reliably answer business questions which have quickly expanded in scale and complexity. To do good data work.