Big Data and Data Analytics

Ajay Mule

Computer Engineering Department, Thakur Polytechnic, Mumbai, India

Abstract: ***Data analytics (DA) is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software. Data analytics technologies and techniques are Widely used in commercial industries to enable organizations to make more-informed business decisions and by scientists and researchers to verify or disprove scientific models, theories and hypotheses.***

Keywords: Data Analytics, Data Analyze, Data Search

1. Introduction

As a term, data analytics predominantly refers to an assortment of applications, from basic business intelligence (BI), reporting and online analytical processing (OLAP) to various forms of advanced analytics. In that sense, it's similar in nature to business analytics, another umbrella term for approaches to analyzing data -- with the difference that the latter is oriented to business uses, while data analytics has a broader focus. The expansive view of the term isn't universal, though: In some cases, people use data analytics specifically to mean advanced analytics, treating BI as a separate category.

Data analytics initiatives can help businesses increase revenues, improve operational efficiency, optimize marketing campaigns and customer service efforts, respond more quickly to emerging market trends and gain a competitive edge over rivals -- all with the ultimate goal of boosting business performance. Depending on the particular application, the data that's analyzed can consist of either historical records or new information that has been processed for real-time analytics uses. In addition, it can come from a mix of internal systems and external data sources.

2. The Dawn of Big Data

Data becomes big data when its volume, velocity, or variety exceeds the abilities of your IT systems to ingest, store, analyze, and process it. Many organizations have the equipment and expertise to handle large quantities of structured data—but with the increasing volume and faster flows of data, they lack the ability to “mine” it and derive actionable intelligence in a timely way. Not only is the volume of this data growing too fast for traditional analytics, but the speed with which it arrives and the variety of data types necessitates new types of data processing and analytic solutions. However, big data doesn’t always fit into neat tables of columns and rows. There are many new data types, both structured and unstructured, that can be processed to yield insight into a business or condition. For example, data from twitter feeds, call detail reports, network data, video cameras, and equipment sensors ften isn’t stored in a data warehouse until you have pre-processed it to distill and summarize and perhaps to detect basic trends and associations. It is more cost effective to load the results into a warehouse for additional analysis. The idea is to “reduce” the data to the point that it can be put in a structured form. Then it can be meaningfully compared to the rest of your data, and scrutinized with traditional business intelligence (BI) tools.

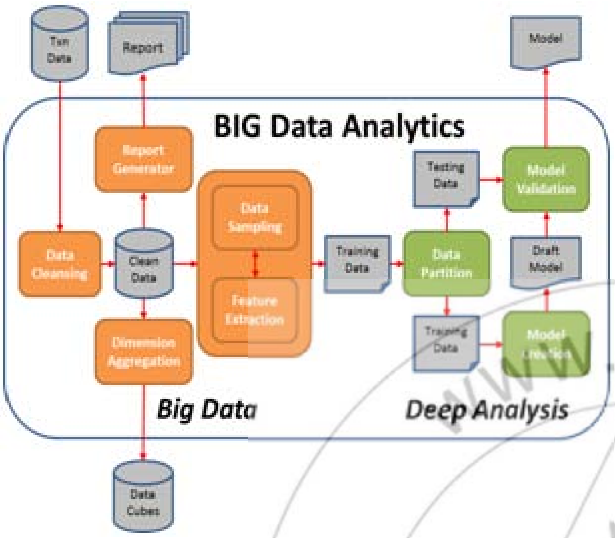
3. Types of Data Analytics Applications

At a high level, data analytics methodologies include exploratory data analysis (EDA), which aims to find patterns and relationships in data, and confirmatory data analysis (CDA), which applies statistical techniques to determine whether hypotheses about a data set are true or false. EDA is often compared to detective work, while CDA is akin to the work of a judge or jury during a court trial. Data analytics can also be separated into quantitative data analysis and qualitative data analysis. The former involves analysis of numerical data with quantifiable variables that can be compared or measured statistically. The qualitative approach is more interpretive -- it focuses on understanding the content of non-numerical data like text, images, audio and video, including common phrases, themes and points of view.

At the application level, BI and reporting provides business executives and other corporate workers with actionable information about key performance indicators, business operations, customers and more. In the past, data queries and reports typically were created for end users by BI developers working in IT or for a centralized BI team; now, organizations increasingly use self-service BI tools that let execs, business analysts and operational workers run their own ad hoc queries and build reports themselves.

4. Process

Data analytics applications involve more than just analyzing data. Particularly on advanced analytics projects, much of the required work takes place upfront, in collecting, integrating and preparing data and then developing, testing and revising analytical models to ensure that they produce accurate results. In addition to data scientists and other data analysts, analytics teams often include data engineers, whose job is to help get data sets ready for analysis.



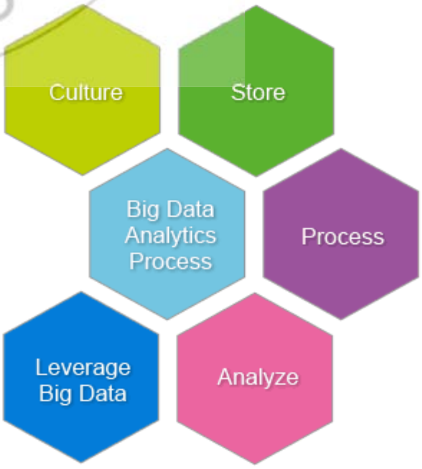
Once the data that's needed is in place, the next step is to find and fix data quality problems that could affect the accuracy of analytics applications. That includes running data profiling and data cleansing jobs to make sure that the information in a data set is consistent and that errors and duplicate entries are eliminated. Additional data preparation work is then done to manipulate and organize the data for the planned analytics use, and data governance policies are applied to ensure that the data hews to corporate standards and is being used properly.

At that point, the data analytics work begins in earnest. A data scientist builds an analytical model, using predictive modeling tools or other analytics software and programming languages such as Python, Scala, R and SQL. The model is initially run against a partial data set to test its accuracy; typically, it's then revised and tested again, a process known as "training" the model that continues until it functions as intended. Finally, the model is run in production mode against the full data set, something that can be done once to address a specific information need or on an ongoing basis as the data is updated.

In some cases, analytics applications can be set to automatically trigger business actions -- for example, stock trades by a financial services firm. Otherwise, the last step in the data analytics process is communicating the results generated by analytical models to business executives and other end users to aid in their decision-making. That usually is done with the help of data visualization techniques, which analytics teams use to create charts and other infographics designed to make their findings easier to understand. Data visualizations often are incorporated into BI dashboard applications that display data on a single screen and can be updated in real time as new information becomes available.

5. Techniques of Analyzing Big Data

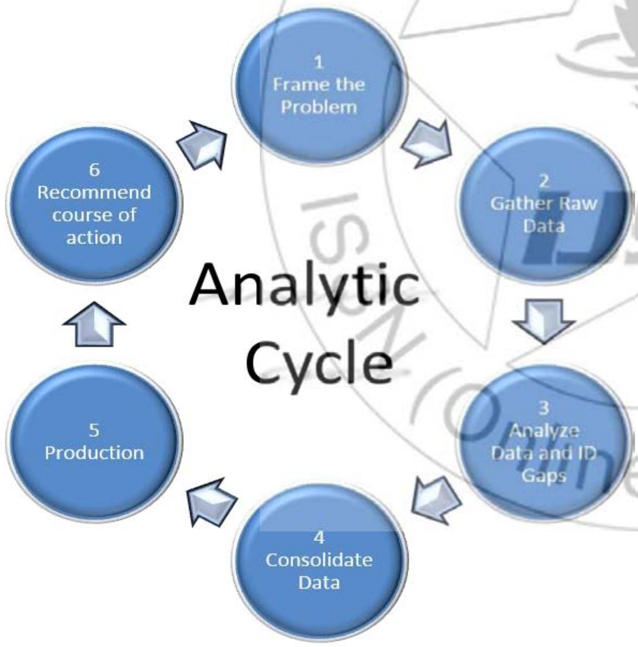
A New Approach When you use SQL queries to look up financial numbers or OLAP tools to generate sales forecasts, you generally know what kind of data you have and what it can tell you. Revenue, geography and time all relate to each other in predictable ways. You don't necessarily know what the answers are but you do know how the various elements of the data set relate to each other. BI users often run standard reports from structured databases that have been carefully modeled to leverage these relationships. Big data analysis involves making —sense" out of large volumes of varied data that in its raw form lacks a data model to define what each element means in the context of the others. There are several new issues you should consider as you embark on this new type of analysis: • Discovery – In many cases you don't really know what you have and how different data sets relate to each other. You must figure it out through a process of exploration and discovery. • Iteration – Because the actual relationships are not always known in advance, uncovering insight is often an iterative process as you find the answers that you seek. The nature of iteration is that it sometimes leads you down a path that turns out to be a dead end. That's okay – experimentation is part of the process. Many analysts and industry experts suggest that you start with small, well-defined projects, learn from each iteration, and gradually move on to the next idea or field of inquiry. • Flexible Capacity – Because of the iterative nature of big data analysis, be prepared to spend more time and utilize more resources to solve problems. • Mining and Predicting – Big data analysis is not black and white. You don't always know how the various data elements relate to each other. As you mine the data to discover patterns and relationships, predictive analytics can yield the insights that you seek. • Decision Management – Consider the transaction volume and velocity. If you are using big data analytics to drive many operational decisions (such as personalizing a web site or prompting call center agents about the habits and activities of consumers) then you need to consider how to automate and optimize the implementation of all those actions. For example you may have no idea whether or not social data sheds light on sales trends.



The challenge comes with figuring out which data elements relate to which other data elements, and in what capacity. The process of discovery not only involves exploring the data to understand how you can use it but also determining how it relates to your traditional enterprise data. New types of inquiry entail not only what happened, but why. For example, a key metric for many companies is customer churn. It's fairly easy to quantify churn. But why does it happen? Studying call data records, customer support inquiries, social media commentary, and other customer feedback can all help explain why customers defect? Similar approaches can be used with other types of data and in other situations. Why did sales fall in a given store? Why do certain patients survive longer than others? The trick is to find the right data, discover the hidden relationships, and analyze it correctly.

6. Advancement in Data Analytic

More advanced types of data analytics include data mining, which involves sorting through large data sets to identify trends, patterns and relationships; predictive analytics, which seeks to predict customer behavior, equipment failures and other future events; and machine learning, an artificial intelligence technique that uses automated algorithms to churn through data sets more quickly than data scientists can do via conventional analytical modeling. Big data analytics applies data mining, predictive analytics and machine learning tools to sets of big data that often contain unstructured and semi-structured data. Text mining provides a means of analyzing documents, emails and other text-based content.



Data analytics initiatives support a wide variety of business uses. For example, banks and credit card companies analyze withdrawal and spending patterns to prevent fraud and identity theft. E-commerce companies and marketing services providers do clickstream analysis to identify website visitors who are more likely to buy a particular product or service based on navigation and page-viewing patterns. Mobile network operators examine customer data to forecast churn so they can take steps to prevent defections to business rivals; to boost customer relationship management efforts, they and other companies also engage in CRM analytics to segment customers for marketing campaigns and equip call center workers with up-to-date information about callers. Healthcare organizations mine patient data to evaluate the effectiveness of treatments for cancer and other diseases.

7. Conclusion

This paper undertook a detailed performance study of three workloads, and found that for those workloads, jobs are often bottlenecked on CPU and not I/O, network performance has little impact on job completion time, and many straggler causes can be identified and fixed. These findings should not be taken as the last word on performance of analytics frameworks: our study focuses on a small set of workloads, and represents only one snapshot in time. As data-analytics frameworks evolve, we expect bottlenecks to evolve as well. As a result, the takeaway from this work should be the importance of instrumenting systems for blocked time analysis, so that researchers and practitioners alike can understand how best to focus performance improvements. Looking forward, we argue that systems should be built with performance understandability as a first-class concern. Obscuring performance factors sometimes seems like a necessary cost of implementing new and more complex optimizations, but inevitably makes understanding how to optimize performance in the future much more difficult.

References

[1] M. K. Aguilera, J. C. Mogul, J. L. Wiener, P. Reynolds, and A. Muthitacharoen. Performance Debugging for Distributed Systems of Black Boxes. In Proc. SOSP, 2003.

[2] G. Ananthanarayanan, M. C.-C. Hung, X. Ren, I. Stoica, A. Wierman, and M. Yu. GRASS: Trimming Stragglers in Approximation Analytics. In Proc. NSDI, 2014.

[3] Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die by Eric Siegel.

[4] Big Data: A Revolution That Will Transform How We Live, Work, and Think by Victor Mayer-Schönberger and Kenneth Cukier.

[5] Analytics in a Big Data World: The Essential Guide to Data Science and its Applications by Bart Baesens (2014)