BIG DATA ANALYTICS II

ISSUE 2, VOLUME 4
OFFICE OF TECHNOLOGY STRATEGIES (TS)

INTRODUCTION

Big data as a concept refers to the high velocity collection of data in large volumes coming from a variety of sources, resulting in different forms of data collected with great speed in great frequency. In an earlier Tech Insight, we introduced the topic of big data, which when used well, contributes to cost reductions, faster and better decision-making, and the offering of innovative products and services. In this second Tech Insight on the topic, we present the key steps and tools commonly used in this form of analytics and address the Department of Veterans Affairs' (VA) actions toward improving and leveraging big data analytics capabilities.

GATHERING DATA

Big data is collected often and from a variety of sources. A few sources of big data include:

- Information gathered from Internet of Things (IoT) devices such as a personal assistants like Amazon's Alexa
- Opendata.gov hosts public data sets from Federal agencies
- Sales transactions at any major store
- Cookies on websites
- Social media contributions on platforms like Twitter and Facebook
- Banking and financial transactions

Collecting data drives the demand for data management and storage. Data management is influenced by research in artificial intelligence and machine learning. Developments in data storage will allow data scientists to draw conclusions from greater sets, creating opportunities to find statistically significant correlations and patterns.

DATA GOVERNANCE & QUALITY

Before analyzing data, it needs to be consistent with data management standards overseen by data governance functions. Data governance includes establishing processes to build and

maintain standards for data quality, including master data management (MDM). MDM is a comprehensive method of enabling an enterprise to link all of its crucial data to one file, called a master file. Once data is deemed reliable, organizations establish an MDM program that standardizes enterprise data practices.

Data quality is an assessment of data's fitness to serve its purpose in a given context. Aspects of data quality include accuracy, completeness, relevance, reliability, and accessibility. Data quality is affected by the way data is entered, stored, and managed. Data quality assurance (DQA) verifies the reliability and effectiveness of the data.

Storing Data

Once high quality data is collected according to enterprise governance standards, big data requires specialized storage that is cost effective, secure, and capable of managing a high data volume while minimizing delays in analysis. Storage solutions include cloud storage (data stored on remote servers accessed from the Internet, or "cloud"), flash storage (data stored on a semi-conductor), and cold storage (less important data kept on slow storage disks). Hybrids of these systems exist, with enterprises storing data based on “temperature.” Hot data remains easily accessible in flash storage, while cold data is stored in cost-effective archives.

Data and its Metadata

Storing data requires managing metadata, which assigns descriptive details about an individual digital asset. When a digital asset is created, so is metadata about its origin, time, date, and format. For instance, when you make a phone call, the data is the conversation, and the call’s metadata is information on the duration, location, and time of the call. It is crucial that enterprises consistently manage and archive data and metadata assets. Following these practices makes digital assets easy to find and distribute, which is key to ensuring the cost effectiveness of big data.

Data Analysis

Once high quality data is collected, processed, and stored, analysis begins. There are many tools and platforms used to process and format data to support analytics. One of the most popular tools is Apache Hadoop, which runs on a framework that allows for the distributed processing of large data sets across clusters of computers. It includes an ecosystem of tools that support multiple analysis scenarios, such as real-time and batch analyses. As free, open-source software, it is adaptable to many analytics purposes, and underpins much of today’s big data analysis.

Industry trends support the use of managed service providers. In the case of Hadoop, it is offered as part of a managed service provided by companies such as Hortonworks. Hortonworks Data Platform (HDP) includes Hadoop and is used for storing, processing, and analyzing large volumes of data. The platform processes data from a variety of sources and in differing formats. HDP is currently available as approved software within the One-VA Technical Reference Model (TRM).

Analytic Approaches

There are three predominant analytics approaches: predictive, batch, and real-time analytics.

Predictive analytics use data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes based on historical data. The goal is to base business decisions on the strongest predictions possible. Common applications of predictive analytics include fraud detection and marketing.

Real-time analytics processes data as it enters an analytics system such as Apache Spark. Real time processing requires a continual input, constant processing, and steady output of data. A few applications of real-time processing are data streaming, radar systems, and bank ATMs. In these cases, immediate processing is crucial to deliver service. Additionally, machine learning provides many ways to support predictive analytics based on data streaming from sources.

Batch analytics processing can take hours or days, and is used for less time-sensitive analysis. This approach involves data collected over a period of time, automated data processing by a separate program, and data output. Google's MapReduce is a common tool for batch processing that both maps and reduces the data. Batch processing is applied in monthly payroll and billing activities.

Visualization

Once processing is complete and insight is gained from the analytics, data visualizations are designed to convey patterns, correlations, and trends in the data. Visualizations help users understand the significance of the data by translating findings into universal, visual language. Programs such as RStudio and Tableau are designed to produce data visualizations.

Harnessing Big Data at VA

VA is adjusting its practices to benefit from big data analytics to support more efficient and effective service delivery to Veterans and beneficiaries. Across VA, meeting emerging data

needs requires new technologies and a new approach to managing analytics. Currently, VA is evaluating IT investments that support big data and planning for future IT investments that support next-generation, cloud-based analytics services, and "data lakes" based on open standards. Also, the VA Office of Information & Technology (OI&T) established an Enterprise Data Management (EDM) function with continued oversight provided by the Data Governance Council (DGC). VA is continuing to evaluate data assets and provide them as Open Data in accordance with Federal policies. Changes underway at VA will also allow the Department to respond to changing technological needs and opportunities, and to fully leverage newly deployed analytic capabilities.

Read more about big data analytics in Enterprise Design Patterns addressing Enterprise Data Analytics and Data Storage, and in Tech Insight Volume 3 Issue 5 on Big Data Analytics. If you have any questions regarding big data analytics, do not hesitate to ask TS for assistance.

TS TECH INSIGHT SERIES

The monthly Tech Insight series aims to help readers make better decisions and be more informed customers (of Office of Information & Technology's products and services) by providing them with high-level overviews of technology issues that impact or will impact VA's Information Technology (IT) environment. Tech Insights introduce topics in an easily digestible fashion by presenting background information on the topic, clearly explaining its importance within VA, and providing recommendations for success from TS. View all TS Tech Insights here.

DISCLAIMER: This document includes links to websites outside VA control and jurisdiction. VA is not responsible for the privacy practices or the content of non-VA websites. We encourage you to review the privacy policy or terms and conditions of those sites to fully understand what information is collected and how it is used.