Introduction to Big Data

Introduction

In the information age, data in huge quantities has become available to analysts and decision-makers. Due to a vast increase in the amount of such data in recent times, a number of specialized platforms and development paradigms have been developed that can handle big data. Using such specialist approaches allows data scientists to gain valuable insights from complex data, ranging from daily transactions to customer interactions and social network data.

This section aims to focus on some of the different analytical approaches and tools data scientists apply to big data in order to gain valuable insights that aid business decision making.

Objectives

You will be able to:

- List the domain areas where big data is particularly useful
- List the technologies associated with big data
- Describe the 3 V's of big data and how they differentiate big data from routine data

What is Big Data

The topic of "big data" has received a lot of hype lately, accompanied by a huge amount of interest from big businesses as it can potentially provide them data-driven decision-making abilities. Big data is one of the most discussed topics in business today across industry sectors, although it was barely known a few years ago. This lesson will focus on what big data is, why it is important, and the benefits it brings.

Big data is no different than normal data that we have seen so far; it's only "bigger." This changes the analytical landscape that must be used as the huge size increase of the data requires specialized tools, techniques, and platforms. It helps us solve new problems and find improved ways to find answers to old problems.

Defining Big Data

Despite all the hype around this topic, there is no clear consensus on how to define **big data**. The term often gets related to business analytics and data mining for identifying relationships and associations present in huge amounts of transaction data.

In the data science domain, big data usually refers to datasets that grow so large that they become awkward to work with using traditional database management systems and analytical approaches. They are datasets whose size is beyond the ability of commonly used software tools and storage systems to capture, store, manage, as well as process the data within a tolerable elapsed time.

How Big is "Big" Data?

Big data sizes are constantly increasing, currently ranging from a few terabytes (TB) to many petabytes (PB) of data in a single dataset. Consequently, some of the difficulties related to big data include capturing, storing, searching, sharing, analyzing, and visualizing. Today, enterprises are exploring large volumes of highly detailed data to discover trends and pieces of information considered incapable of being captured before.

Here are some of the examples of big data:

- Web traffic data: Data points such as number of page views, previous web page, user information, advertisement click-through rate, pages per visit, average visit duration
- Text data: Emails, tweets, news reports, voice recordings, and text gathered from crawling the web can make massive datasets that are valuable to data scientists
- Location and time data: GPS data helps Google determine which roads have higher traffic and which businesses will be busier at certain hours
- Social network data: Using the information of relationships between users on Facebook, LinkedIn, Twitter, Reddit, and countless other websites and apps
- Smart grid and sensor data: With the advent of the Internet of Things (IoT), more and more devices are able to record data at all times, making it possible to gather lots of data instantaneously

3 V's of Big Data

Doug Laney published a **paper (https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf)** on three defining characteristics of big data. Three main features characterize big data: volume, variety, and velocity, or the three V’s. The volume of the data is its size, and how enormous it is. Velocity refers to the rate with which data is changing, or how often it is created. Finally, variety includes the different formats and types of data, as well as the different kinds of uses and ways of analyzing the data:

Let's look a bit deeper into what these 3 V's refer to:

VOLUME

Volume refers to the **amount of data** generated through websites, portals, and online applications in a data-driven business. Especially for online retailers, volume encompasses the available data that are out there and need to be assessed for relevance.

Consider the following:

As of 2019, Facebook has 2.32 billion users, Youtube: 1.9 billion users, WhatsApp: 1.6 billion users and Instagram: 1 billion users. Every day, these users contribute to billions of images, posts, videos, tweets, etc. You can now imagine the insanely large amount (or **v**olume) of data that is generated every minute around the world. Data volume is the primary attribute of big data. Big data can be quantified by size in Terabytes (TBs) or Petabytes (PBs), as well as even the number of records, transactions, tables, or files. Additionally, one of the things that makes big data really big is that it's coming from a greater variety of sources than ever before, including logs, clickstreams, and social media as we will see below.

VELOCITY

Velocity refers to the speed with which data is generated, and as internet speeds have increased and the number of users has increased, the velocity has also increased substantially.

The following image created by **Lori Lewis and Chadd Callahan (https://www.allaccess.com/merge/archive/29580/2019-this-is-what-happens-in-an-internet-minute)** shows what happens on major social media platforms in one minute.

Velocity is basically the frequency of data generation or the frequency of data delivery. The leading edge of big data is streaming data, which is collected in real-time from the websites.

Tools within the big data stack help companies hold this explosion in velocity, accept the incoming flow of data, and at the same time process it quickly enough so that it does not create bottlenecks.

VARIETY

Variety in big data refers to all the structured and unstructured data that has the possibility of getting generated either by humans or by machines. Structured data is whatever data you could store in a spreadsheet. It can easily be cataloged and summary statistics can be calculated for it. Unstructured data are raw things like texts, tweets, pictures, videos, emails, voice mails, hand-written text, ECG readings, and audio recordings. Humans can only make sense of data that is structured, and it is usually up to data scientists to create some organization and structure to unstructured data.

Variety is all about the ability to classify the incoming data into various categories and turn unstructured data into something with more structure.

This leads us to the most widely used definition in the industry by Gartner:

***Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation*.**

Any data sources that fall under those 3 Vs are sources of big data, no matter how you define it.

NOTE: *Some researchers have discussed the addition of a fourth V, or **Veracity**. Veracity focuses on the quality of the data. This characterizes big data quality as good, bad, or undefined due to data inconsistency, incompleteness, ambiguity, latency, deception, and approximations*

The important thing to remember from this three-pronged definition of Big Data is that there is not a single component that makes data "big" or not. It is also a futile effort to try and make a definition of what the threshold is to make something "big data" rather than normal data. As technologies evolve and new distributed algorithms are created, what was once big data will no longer be big, and we will raise the bar.

Big Data Analytics

With the evolution of technology and the increased amounts of data, as discussed above, the need for faster and more efficient ways of analyzing such data has also grown exponentially. Having big data **alone** is no longer enough to make efficient decisions at the right time. As we mentioned above, Big Data cannot be easily analyzed with traditional data management and analysis techniques and infrastructures. Therefore, there arises a need for new tools and methods specialized for big data analytics, as well as the required architectures for storing and managing such data. Accordingly, the emergence of big data has an effect on everything from the data itself to its collection, analysis, and visualization, as well as the final extracted decisions.

The image below shows the technology stack, or the key tools and platforms being heavily employed in big data analytics today.

Explaining each one of these tools/platforms etc. is outside the scope of this lesson. You are, however, encouraged to look up these technologies and see their role in big data analytics. Such a stack maps the different big data storage, management, analytics tools/methods, visualization, and evaluation tools to the different phases of the decision-making process.

The key activities associated with big data analytics are reflected in four main areas:

- Big data warehousing and distribution
- Big data storage
- Big data computational platforms
- Big data analyses, visualization, and evaluation

Such a framework can be applied for knowledge discovery and informed decision-making in big data-driven organizations.

Example Business Applications of Big Data Analytics

Along with some of the most common advanced data analytics methods such as regression analysis, association rules, clustering, and classification, some additional analyses have become common with big data.

For example, social media has recently become important for social networking and content sharing. Yet, the content that is generated from social media websites is enormous and remains largely unexploited. However, social media analytics can be used to analyze such data and extract useful information and predictions.

> **Social media analytics** is based on developing and evaluating informatics frameworks and tools in order to collect, monitor, summarize, analyze, as well as visualize social media data.

Social media analytics facilitates understanding the reactions and conversations between people in online communities, as well as extracting useful patterns and intelligence from their interactions and what they share on social media websites.

On the other hand, **text mining** and **NLP** techniques are used to analyze a document or set of documents in order to understand the content within and the meaning of the information contained. Text mining has become very important nowadays since much of the information stored consists of text - in the form of emails, SMS texts, social media feeds, blogs, etc. While data mining deals with structured data, text presents special characteristics which basically follow a non-relational form and require wisely thought-out schemas to grant it more structure.

Sentiment analysis/opinion mining is also becoming more and more important as online opinion data, such as blogs, product reviews, forums, and social data from social media sites, like Twitter and Facebook, grow tremendously.

> **Sentiment Analysis** focuses on analyzing and understanding emotions from subjective text patterns and is enabled through text mining. It identifies the opinions and attitudes of individuals towards certain topics, and it is useful in classifying viewpoints as positive or negative.

Sentiment analysis uses NLP and text analytics in order to identify and extract information by finding words that are indicative of certain sentiments, as well as relationships between words so that sentiments can be accurately identified.

And finally, one of the leading applications in big data analytics is **recommendation systems**. Powerful recommendation engines can be built for anything from movies and videos to music, books, and products as offered by Netflix, Pandora, or Amazon. As customers of an online retailer browse through products, the Recommendation system offers recommendations of products they might be interested in. In our daily online browsing and shopping routine, most of us often come across messages like the one shown below. This is a recommendation system doing its job.

Recommendation systems have been immensely beneficial for both businesses and consumers. Big data is the driving force behind recommendation systems. A typical recommendation system cannot do its job without sufficient data and big data supplies plenty of user data such as past purchases, browsing history, and feedback for the recommendation systems to provide relevant and effective recommendations. In a nutshell, even the most advanced recommendations cannot be effective without big data.

So what's next?

After this quick introduction, we will look at MapReduce, a distributed computation platform designed to incorporate big data analytics and how it is used by Hadoop/Apache Spark development environments to analyze big data.

Additional Reading

Big data is a huge subject and incorporates a lot of underlying technologies and principles. You are advised to visit the following resources and read up on big data to develop a sound and holistic understanding of the domain.

- **Youtube: Big Data Trap (https://www.youtube.com/watch?v=0cizsKDn3TI)** (https://www.youtube.com/watch?v=0cizsKDn3TI) - Highly recommended, an excellent lecture on the social dimension of big data.

- **Big Data vs Data Science** (https://www.educba.com/big-data-vs-data-science/) - How to relate big data analytics to routine analytics that we have so far!
- **Big Data Analytics** (https://pdfs.semanticscholar.org/d392/0f02dbb15da19b04d782fc0546ef113e0bf7.pdf) - A great paper summarizing big-data-related terms, ideas, etc.
- **Introduction to Big Data** (https://web.archive.org/web/20200214174508/https://www.ntnu.no/iie/fag/big/lessons/lesson2.pdf) - A paper discussing the basics of big data

Summary

In this introductory lesson on big data, we looked at what data qualifies as "big data". We looked at how it is hard to come up with a standard definition of big data due to the variety of its applications and use cases. Up next, we will get into how we actually make parallelizable applications that are efficient with big data.