Kunnskap for en bedre verden

Introduction to Big Data

Xiaomeng Su, Institutt for informatikk og e-læring ved NTNU
Learning material is developed for course IINI3012 Big Data

Summary: This chapter gives an overview of the field big data analytics. We start with defining the term big data and explaining why it matters. We then move on to give some examples of the application area of big data analytics. The people who work on big data analytics are called data scientist these days and we explain what it encompasses. Finally, we outline the main technological components in a big data environment.

1. What is Big Data and why does it matter?

It is difficult to recall a topic that received so much hype as broadly and as quickly as big data. While barely known a few years ago, big data is one of the most discussed topics in business today across industry sectors. This section has focus on what big data is, why it is important, and the benefits of analysing it.

1.1. What is big data analytics?

As one of the most "hyped" terms in the market today, there is no consensus as to how to define big data. The term is often used synonymously with related concept such as *Business Intelligence* (BI) and *data mining*. It is true that all three terms is about analyzing data and in many cases advanced analytics . But big data concept is different from the two others when data volumes, number of transactions and the number of data sources are so big and complex that they require special methods and technologies in order to draw insight out of data (for instance, traditional data warehouse solutions may fall short when dealing with big data).

This also forms the basis for the most used definition of big data, the three V: *Volume*, *Velocity* and *Variety* as shown in Figure 1.

- Volume: Large amounts of data , from datasets with sizes of terabytes to zettabyte.
- Velocity: Large amounts of data from transactions with high refresh rate resulting in data streams coming at great speed and the time to act on the basis of these data streams will often be very short . There is a shift from batch processing to real time streaming.
- Variety: Data come from different data sources. For the first, data can come from both internal and external data source. More importantly, data can come in various format such as transaction and log data from various applications , structured data as database table , semi-structured data such as XML data, unstructured data such as text, images, video streams, audio statement, and more. There is a shift from sole structured data to increasingly more unstructured data or the combination of the two.

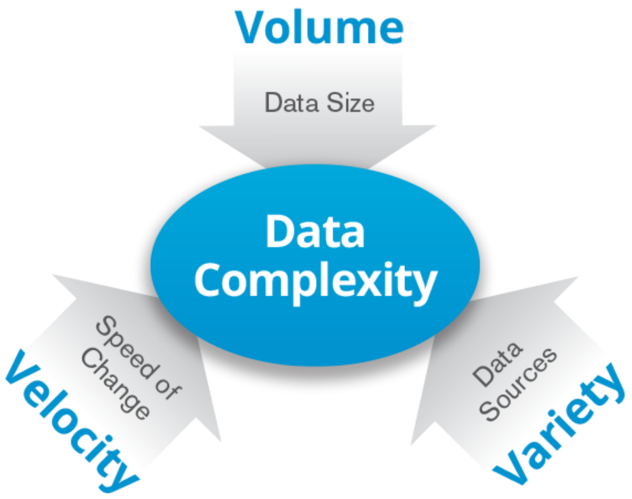


Figure 1 The three V of Big Data

This leads us to the most widely used definition in the industry. Gartner (2012) defines Big Data in the following.

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.

It should by now be clear that the "big" in big data is not just about volume. While big data certainly involves having a lot of data, big data does not refer to data volume alone. What it means is that you are not only getting a lot of data. It is also coming at you fast, it is coming at you in complex format, and it is coming at you from a variety of sources.

It is also important to point out that there might not be too much value in defining an absolute threshold for what constitutes *big* data. Today's big data may not be tomorrow's big data as technologies evolve. It is, by and large, a relative concept. From anyone's given perspective, if your organization is facing significant challenges (and opportunities) around data's volume, velocity and variety, it is your big data challenge. Typically, these challenges introduce the need for distinct data management and delivery technologies and techniques.

1.2. What data are we talking about?

Organizations have a long tradition of capturing transactional data. Apart from that, organizations nowadays are capturing additional data from its operational environment at an increasingly fast speed. Some example are listed here.

- Web data. Customer level web behaviour data such as page views, searches, reading reviews, purchasing, can be captured. They can enhance performance in areas such as next best offer, churn modelling, customer segmentation and targeted advertisement.
- Text data (email, news, Facebook feeds, documents, etc) is one of the biggest and most widely applicable types of big data. The focus is typically on extracting key facts from the text and then use the facts as inputs to other analytic process (for example, automatically classify insurance claims as fraudulent or not.)
- Time and location data. GPS and mobile phone as well as Wi-Fi connection makes time and location information a growing source of data. At an individual level, many organizations come to realize the power of knowing when their customers are at which location. Equally important is to look at time and location data at an aggregated level. As more individuals open up their time and location data more publicly, lots of interesting applications start to emerge. Time and location data is one of the most privacy-sensitive types of big data and should be treated with great caution.
- Smart grid and sensor data. Sensor data are collected nowadays from cars, oil pipes, windmill turbines, and they are collected in extremely high frequency. Sensor data provides powerful information on the performance of engines and machinery. It enables diagnosis of problems more easily and faster development of mitigation procedures.
- Social network data. Within social network sites like Facebook, LinkedIn, Instagram, it is possible to do link analysis to uncover the network of a given user. Social network analysis can give insights into what advertisements might appeal to given users. This is done by considering not only interests the customers have personally

> stated, but also knowing what it is that their circle of friends or colleagues has an interest in.

With most of the big data source, the power is not just in what that particular source of data can tell you uniquely by itself. The value is in what it can tell you in combination with other data (for instance, a traditional churn model based on historical transaction data can be enhanced when combined with web browsing data from customers.). It really is the combination that counts.

1.3. How is big data different from traditional data sources?

There are some important ways that big data is different from traditional data sources. In his book *Taming the big data tidal wave,* the author Bill Franks suggested the following ways where big data can be seen as different from traditional data sources.

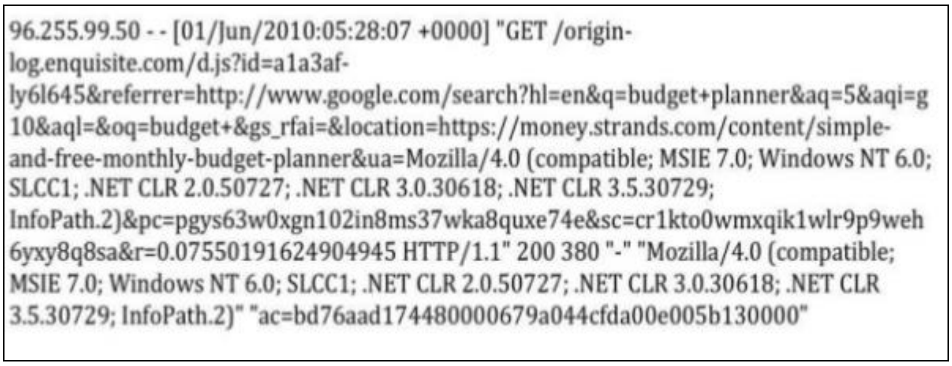
First, big data can be an entirely new source of data. For example, most of us have experience with online shopping. The transactions we execute are not fundamentally different transactions from what we would have done traditionally. An organization may capture web transactions, but they are really just more of the same transactions that have been captured for years (e.g. purchasing records). However, actually capturing browsing behaviour (how do you navigate on the site, for instance) as customers execute a transaction creates fundamentally new data.

Second, sometimes one can argue that the speed of data feed has increase to such an extent that it qualifies as a new data source. For example, your power meter has probably been read manually each month for years. Now we have a smart meter that automatically read it every 10 minutes. One are argue that it is the same data. It can also be argued that the frequency is so high now that it enables a very different, more in-depth level of analytics that such data is really a new data source.

Third, increasingly more semi-structured and unstructured data are coming in. Most traditional data sources are in the structured realm. Structure data are the ones like the receipts from your grocery store, the data on your salary slip, accounting information on the spreadsheet, and pretty much everything that can fit nicely in a relational database. Every piece of information included is known ahead of time, comes in a specified format and occurs in a specified order. This makes it easy to work with.

Unstructured data sources are those that you have little or no control over its format. Text data, video data and audio data all fall into this category. Unstructured data is messy to work with because the meaning of the bites and bits are not predefined.

In between structured and unstructured data is semi-structured data. Semi-structured data is data that may be irregular or incomplete and have a structure that may change rapidly or unpredictably. It generally has some structure, but does not conform to a fixed schema. Web logs are good example of semi-structured data. See Figure 2 for an example of a raw web log. Web logs look messy. However, each piece of information does, in fact, serve a purpose of some sort. For example, in Figure 2, *referrer = http://www.google.com/search* tells us what is the referral channel (i.e. in this case the user landed on this web page through google search) . The log text generated by a click on a website right now can be longer or shorter than the log text generated by a click from a different page a minute later. In the end, however, it is important to understand that semi-structured data does have an underlying logic. It simply takes more effort (with the help of natural language processing tools) than structured data to develop relationships between various pieces of it.



```
96.255.99.50 - - [01/Jun/2010:05:28:07 +0000] "GET /origin-
log.enquisite.com/d.js?id=a1a3af-
ly6l645&referrer=http://www.google.com/search?hl=en&q=budget+planner&aq=5&aqi=g
10&aql=&oq=budget+&gs_rfai=&location=https://money.strands.com/content/simple-
and-free-monthly-budget-planner&ua=Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0;
SLCC1; .NET CLR 2.0.50727; .NET CLR 3.0.30618; .NET CLR 3.5.30729;
InfoPath.2)&pc=pgys63w0xgn102in8ms37wka8quxe74e&sc=cr1kto0wmxqik1wlr9p9weh
6yxy8q8sa&r=0.07550191624904945 HTTP/1.1" 200 380 "-" "Mozilla/4.0 (compatible;
MSIE 7.0; Windows NT 6.0; SLCC1; .NET CLR 2.0.50727; .NET CLR 3.0.30618; .NET CLR
3.5.30729; InfoPath.2)" "ac=bd76aad17448000679a044cfda00e005b130000"
```

Figure 2 Example of a raw web log (from http://www.decideo.fr/bruley/).

Is it more important to work with big data than with traditional data? Reading a lot of hype around big data, one may start to think that just because big data has high volume, velocity and variety, it is somehow better or more important than other data. This is not the case. The power of big data is in the analysis you do with it and the actions you take as the result of the analysis. Big data or small data does not in and by itself possession any value. It is valuable only when you can get some insight out of the data. And that insight can be used to guild your decision making.

1.4. Different level of "insight" – from descriptive to predictive and prescriptive

Along with big data, there is also a so-called paradigm shift in terms of analytic focus. That is a shift from descriptive analytics to predictive and prescriptive analytics.

Descriptive analytics answers the questions about "what happened in the past?" This involves typically reporting. We can look at some example questions that are typically addressed here.

- What was the sales revenue in the first quarter of the year? Is additional sales effort needed to meet our target?
- Which is our most profitable product/region/customer?
- How many customers did we win/loose in the first half-year? How many did we win/loose in Oslo area, how many in Mid Norway?
- How many of the won customers can be attributed to the promotional campaign (e.g. via a recorded promotional code) that was launched in Mid Norway last month? Was the campaign successful?

Predictive analytics aim to something about "what might happen next?" This is harder and it involves extrapolating trends and patterns to the future. Some example questions look like this.

- What will the number of complaints to our call centre next quarter?

- Which customer are most likely to churn (e.g. cancel her subscription)?
- What is the next best offer for this customer?

Prescriptive analytics tries to answer, "how do I deal with this". This is where analytics gets operational. It is totally business and use case dependent. Some examples to illustrate the point.

- We know that this person has a high chance to churn, we can offer her a value package.
- We know the viewing history of this customer on our news site, we can recommend articles that we think she would like to read next.
- From analysing various sensor data we know that part A of windmill 101 is about to break, a replacement part is automatic ordered through supply chain.

All three type of analytics existed before the big data era but the focus has traditionally been on reporting. The difference that big data brought to the table are twofold: i) the appetite and ability for precise forward-looking insight ii) the appetite and ability for fast and actionable insight. Forward-looking insights means that business now has the appetite and ability to predict what might happen next. Traditionally, we can also do that, but the accuracy was far less impressive given the limited amount and source of data. Big data change this equation. Fast and actionable insight means that whatever we get out of the data analysis has to have an impact on the business process and preferable the impact is embedded in the process. For instance, recommender systems automatically generate personalized recommendations (e.g. what Amazon recommends you to buy is different than what it recommends me to buy since we have different purchasing and viewing history) right after a purchasing transaction in the hope to increase sales there and then.

This is not to say that descriptive analytics is not important. Reporting has been and will still be an important part of business life. In practice, one should not be rigid and insisting on only one or another type of analytics. What yields most benefit is of course depending on the nature of the business question and thereafter choosing "the right tool for the right job".

1.5. The business value of big data analytics

Let us revisit the definition from 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****.*

This definition from Gartner summarized succinctly the main benefits of big data analytics. They are i) to draw insight from data, ii) to make better decision based on the insight, and iii) to automate the decision and bake it into a business process, hence process automation.

In a more detailed level, each big data solution may address particular business problems the organizations face and the business value of the solution is further connected to the original business problems. For instance, churn prediction will hopefully reduce churn and therefore avoid declining of revenue. When building a business case for big data analytics project, it is important to start with a business problem, not data or technology. Collecting data or purchasing technology without a clear business target is a losing strategy. A business case for analytics must solve real business problems that an organization faces.

2. Applications of big data across industry sectors

We will describe and discuss big data application in more detail in later chapters. Here we introduce a few concrete examples to make the understanding more grounded.

2.1. Segmentation and prediction

A large amount of big data applications fall in the cateory of classification and prediction. Take banks for an example[1]. Every day millions of people apply for new credit cards, loans, and mortgages. In the decision-making process, banks use one number to review a person's financial history and assess their likelihood to pay off debt: a credit score. This score is calculated from all the data the banks knows about you. Similarly, other industries are trying to mimic this approach by using algorithm-based data to predict future outcomes in various settings. Take for instance the trend of wearing devices to measure biometrics, such as fitness activity, sleep patterns, and calorie intake. With the ability to monitor these types of metrics, physicians and health insurance companies will have the ability to predict health outcomes and behaviors.

2.2. Churn prediction

In the telecom sector, customer switch from one company to another is called *churn*. Since attracting new customer is much more expensive than retaining new ones, companies have invested large amounts of time and effort to create and enhance churn model. The intention is to flag customer that at the risk of churning, and find ways to retain them (for instance, by ways of retention incentives) before they leave. Churn is a major issue for the industry and there are huge amounts of money at stake. The churn models have a major impact on the bottom line. Churn models have traditionally been relying on historical data to try to capture the characteristics of those who churned (for instance, usage dropping, particular demographics). And then examine the current user group against these characteristics. Those who are highly similar with the historical "churners" will be flagged and be followed up by sales agent. Now imagine, if the company also have web data from the users and they captured that a user has checked the cancellation policy page of the company (let us forget about identifying users across channel and privacy issues for a moment). This web data can be used to enhance the churn model. In addition, telecom companies are also experimenting with publically available social media data to improve their churn model.

2.3. Recommender systems and targeted marketing

Recommender system are common in almost every domain. They are used for book recommendations on Amazon.com ("customer who bought this item also bought…"), for music recommendations on Spotify, on movie recommendations on Netflix, and on news recommendations on almost all news portal. Some recommendation are based on general trends (e.g. "most read news for today…"), while others are more personalized recommendations (for example, "recommended to you because you have watched…" on

[1] Example taken from Bill Franks *Taming the big data tidal wave*. Readers that are interested in learning more on big data analytics (with focus on business side) from a practitioner's point of view is encouraged to read the book.

Netflix). Most of the Norwegian customer are also familiar with the member broche from Coop where the coupons are personalized offers. Recommender system, when implemented properly can affect business significantly. For instance, Netflix reported that 2/3 of the movies watched are recommended, Google News stated that recommendations generate 38% more click-throughs and Amazon claimed that 35% sales come from recommendations.

2.4. Sentiment analysis

One popular use of text data today is the so called sentiment analysis. Sentiment analysis looks at the general direction of opinions across a large number of people to provide information on what the market is saying, thinking, and feeling about an organization. It often uses data from social media sites as well as other customer touch point. Examples include: what is the buzz around a company or product? Are people saying good or bad things about an organization and the services it offers? Getting a feeling on the treads of what people are saying across social media outlets or within customer service interactions can be valuable in planning what to do next. It can also be used at an individual level. Sentiment analysis can use pattern recognition to detect a caller's mood at the start of a call. An agitated caller might be quickly routed to a specialist for careful treatment.

2.5. Operational analytics

Operational analytics is about embedding analytics within business processes and automation decisions so that millions of decisions every day are made by analytics processes without any human intervention. For instance, airlines automatically reroute customers when a flight is delayed in order to limit travel disruption and raise cutstomer satisfaction. The analytics take into account a lot of facts about each customer, other passengers, and the status of alternative flight options.

2.6. Big data for social good

Insights from refined data can help the business, but it can also foster social good and empower societies. For instance, in 2015, Telenor Research published a study in conjunction with the Harvard T.H. Chan School of Public Health and Telenor Pakistan, demonstrating the power of mobile data to predict and track the spread of epidemic disease. The research received widespread acknowledgement, and was even tweeted (https://twitter.com/billgates/status/649571076163391488) about by Bill Gates.

3. Data scientist

The people that does the big data analytic job are called data scientist nowadays. Thomas H. Davenport and D.J. Patil coined the term "data scientist " in a Harvard Business Review article in 2012. The article described the role and called it "the sexiest job of the 21st century."

The job title "data scientist" is sometimes criticized because it lacks specificity and can be perceived as a glorified synonym for data analyst. Regardless, the position is gaining acceptance with large enterprises who are interested in deriving meaning from big data, the large amount of structured, unstructured and semi-structured data that a large enterprise produces. The primary distinction found in practice between data scientist and other analytics professional is that data scientist are likely to come from a computer science background; to use Hadoop, and to code in languages like Python or R. This compares to traditional analytics professionals who are likely to come from statistics, math, or operations research background and are likely to use relational and analytics server environments to code in SAS and SQL. However, it is not the environment and tool set that fundamentally defines a job. It is the type of business problems that are solved and the core skill set needed. In that sense, data scientist are not so much different from traditional analytics professional where the analytical mind set remains unaltered.

4. The main technological components in a Big Data ecosystem

Let us get back to the Gartner definition for yet another time. *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. It spells out explicitly that big data necessitates a new type of data management solution because of its high-volume, high-velocity and/or high-variety nature. This new type of data management solution bears the trademark of highly scalable, massively parallel, and cost-effective.

4.1. Technologies for capturing, storing and accessing big data

Traditionally, data are stored in relational database (for example a CRM system for customer data, a supply chain management software for vendor related information) and some of these data are extracted periodically from the operational database, transformed and loaded into data warehouse for reporting and further analysis. This is typically in the realm of Business Intelligence. Such process and tool set fall short when dealing with big data. For instance, one of the largest publicly discussed Hadoop cluster (Yahoo’s) was at 455 petabytes in 2014 and it's grown since then. There simply is no parallel relational databases or data warehouse that have come even close to those kinds of numbers. Another sweet spot for Hadoop (over relational technology) is when data comes in unstructured format, such as audio, video, text.

It is worthwhile to mention that there is a general misconception that new technology, such as Hadoop is replacing other technologies, such as relational database. It is not the case. It is more likely that they are being added alongside each other. The sweet spot for a massively parallel relational platform for instance, is dealing with high-value transactional data that is already structured, that needs to support a large amount of user and applications that ask repeated questions of known data (where a fixed schema and optimization pays off) with enterprise level security and performance guarantee.

It is often called the Hadoop eco-system when discussing the various lays of technologies used to deal with big data. For a complete list, please refer to https://hadoopecosystemtable.github.io/. An example stack might look like that.

- Amazon web service for infrastructure (in the Cloud and pay as you go)
- Apache HDFS (Hadoop Distributed File System) for distributed file system
- MapReduce or Spark for distributed programming model
- Cassandra or HBase for non-relational distributed database management system

- Hive for execute SQL on top of Hadoop
- Mahout for Machine learning library and math library, on top of MapReduce.
- R for data analytics and visualization

We will discuss more of the technical elements in later chapters.

4.2. Analytical techniques

Most of the widely used analytical techniques falls into one of the following categories.

- Statistical methods, forecasting, regression analysis
- Database querying
- Data warehouse
- Machine learning and data mining

Later lectures will discuss some of them in greater detail, where methods will be described at a non-technical level, focusing on idea behind method, how it is used, advantages and limitations, and when the method is likely to be of value to which business objective.

4.3. Visualization

When analysis is done, the results need to be communicated to various stakeholders. One of the hardest parts of an analysis is producing quality supporting graphics. Conversely, a good graph is one of the best ways to present findings. Graphics are used primarily for two reasons: exploratory data analysis and presenting results. More on visualization will be introduced in a later chapter.

5. Summary

- Big data is here and it is here to stay. Despite the hype, big data does offer tangible business benefit to organizations. It enables *enhanced insight, decision making, and process automation.*
- The characteristics of big data is the three V: *Volume*, *Velocity* and *Variety*. The “big” in big data is not just about volume. While big data certainly involves having a lot of data, big data does not refer to data volume alone. What it means is that you are not only getting a lot of data. It is also coming at you fast, it is coming at you in complex format, and it is coming at you from a variety of sources.
- Data comes from variety of sources, and can be used in various industry applications. Often it is the combination of data sources that counts.
- Along with big data, there is also a so-called paradigm shift in terms of analytic focus. That is a shift from descriptive analytics to predictive and prescriptive analytics.
- Big data necessitates a new type of data management solution because of its high-volume, high-velocity and/or high-variety nature. This new type of data management solution bears the trademark of highly scalable, massively parallel, and cost-effective.

- New technologies, such as Hadoop, are not replacing other technologies, such as relational database, but rather are being added alongside them.

6. Review questions

1. Think of one example where big data can make an impact, preferably one that is related to your work experience. Describe the example in the following way. What industry sector is it? What is the business problem? How do you think big data can help? What is the business value of such a big data solution? What kind of data source will be used? What kind of analytical techniques would be most suitable in this case?
2. Some companies move their relational database (for instance a reporting database residing in MS SQL server) to Hadoop ecosystem. Do some web search and find out why do companies do that? What do they gain from that? Is it wise to do so? What is the upside and what is the downside?
3. Visit a job-posting web site such as Finn.no. Spend some time at the site examining jobs for data scientist, data analyst, and data science consultant. Find two or three description of jobs and study them. What kind of knowledge do these jobs require? What do you need to do to prepare for these jobs? Write a one page report summarizing your findings.

7. Additional reading

There is one additional reading for Norwegian readers, *"Kartlegging og vurdering av stordata i offentlig sektor"*. It is available on It's learning. The primary intention is to get familiar with big data terminologies in Norwegian language as well as to gain overview of applications of big data in the Norwegian context.

8. References

Mark A. Beyer and Douglas Laney. *"The Importance of 'Big Data': A Definition"*. Gartner, 2012

Bill Franks. *"Taming the big data tidal wave"*. Wiley, 2012

David R. Hardoon and Galit Shmueli. *"Getting started with business analytics – insightful decision making"*. Talor & Francis Group.2013

Foster Provost and Tom Fawcett. *"Data science for busines*s". O'Relly, 2013

Thomas H. Davenport and D.J. Patil . *"Data Scientist: The Sexiest Job of the 21st Century"*, Harvard Business Review, 2012