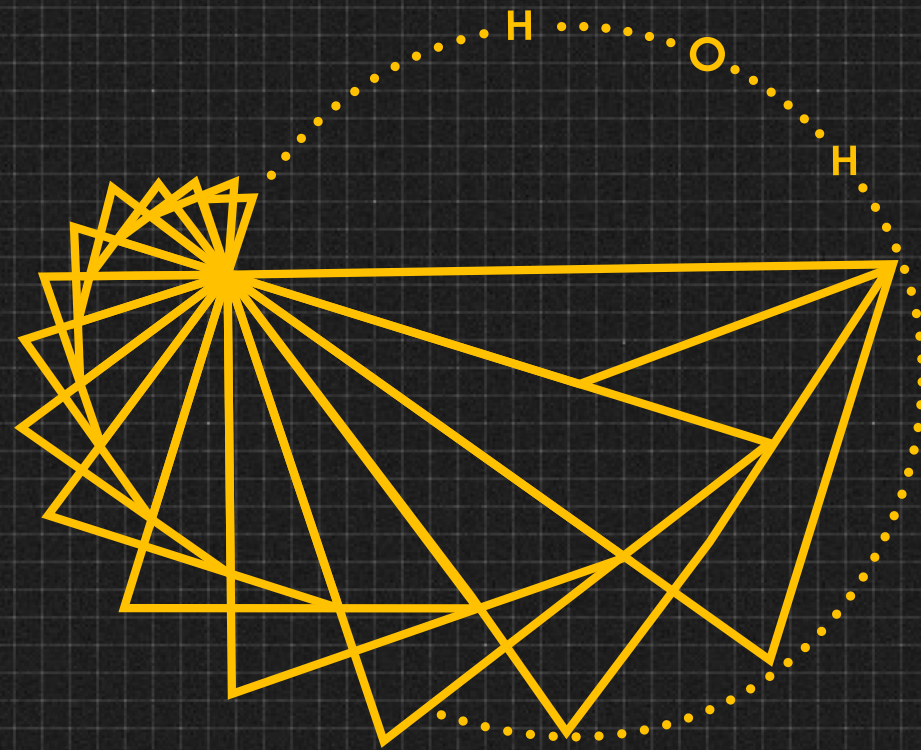




data
iku


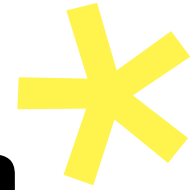
GUIDEBOOK



From Small to Big Data, Adopting the Advanced Analytics Mindset

WHEN SIZE REALLY DOES MATTER

Introduction



So here's the deal. Big data is first and foremost a question of data volume. Instead of collecting comparatively small amounts of data (megabytes, gigabytes), companies are collecting thousands, millions, billions or many, many times more. But “Big” also means combining data from all available sources to get a complete understanding of your business at the finest level of detail. From every customer click on your site to every movement of a machine in your factory, data is no longer siloed between departments and brings insights to everyone.

All this data means finding new ways to process it. Setting up ETL pipelines with your data engineers for every one-off business request can take up months. And Excel can handle a million lines, but it's limited by your computer's desktop RAM.

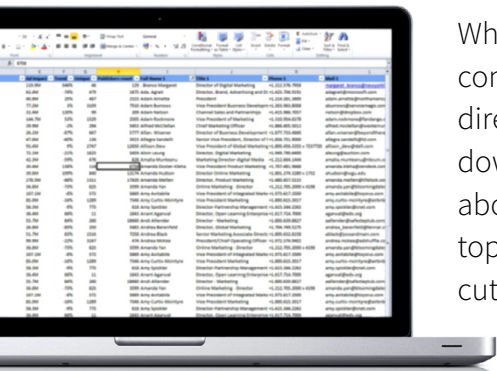
The solution? Going straight to your data source! Instead of exporting and updating, connect directly to your company's database. There are so many tools today that allow you to do so without code, even on Hadoop. You don't need to rely on data scientists to go straight to the data and ask your questions.

However, before you get started, you need to understand the fundamental differences between the small data world, and the brave new big data mindset!



Distributed Processing vs Extracting and Uploading Data: Power Unleashed

As data volumes grow, you want to move the data less and use the power of the data source more. Why? It's simply faster and gives you more power.



While small data can be processed locally (meaning on your own personal computer); processing “big” data is far more demanding. By connecting directly to a database, your calculations and visualisations are pushed down so they’re performed in the database. Take a moment to think about it; big data is stored on rows and rows of hard-drives stacked on top of each other - usually in massive facilities. Then think about how cute your PC is in comparison.

In addition to the extra computing and storage power, connecting directly to the source means that you always have access to the most recent data, meaning that the days of outdated Excel extracts are over.



Key Takeaway

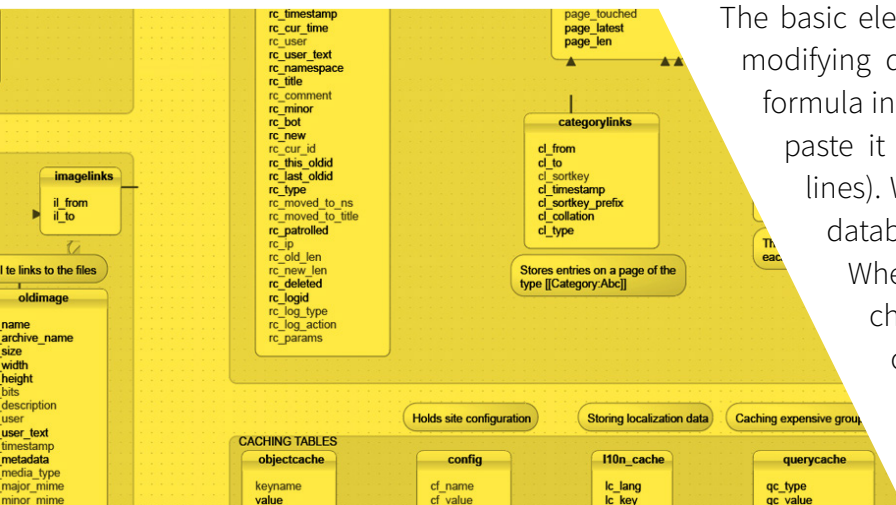
In practice, this means instead of opening your .xls or .csv files in Excel, you set up a connection and interact with your database directly. How do you do this? The most obvious way is by using code (SQL, R, and Python are the most common), but there are also tools to avoid code that work extremely well. You can check out Tableau for data visualization, various ETLs for data preparation, and Dataiku DSS for collaborative in-database analytics.





Mass Actions vs Cell-To-Cell Modifications: Enforce Coherence

This is something of a tricky concept but hang in there - this is a very important idea that will stay with you throughout your data preparation and analysis work.



The basic element you're working on when you're modifying data in Excel is the cell. You write a formula in a cell, and then, if you want, you copy paste it to the rest of your observations (or lines). When you're working with data from a database, your basic element is a column. Whether you're cleaning your data, changing a format, creating categories, or enriching it with new variables, you'll be creating new columns in new datasets; never changing one line of a file at a time.



Key Takeaway

This is good news for you. You won't have to do that extra step of copy pasting your formulas and dealing with \$\$s. No more errors from incomplete column copy-pastings, or from messy filtering!

However, this means that you'll be designing your changes before actually executing them (computations on distributed systems can take hours or even days!). Most tools will allow you to work on smart samples of your data, allowing you to confirm that you're satisfied with the results, before deploying that design and running it on the whole dataset.

Why do it this way? Simple. So the information you store always stays consistent. Every line contains the same types of information in the same formats.



File System vs Complex Workflows:

Hello Automation



Working with your database directly is also going to change your workflow, and how you manage different steps of your process. Let's discuss the global flow of your data analysis projects.

If you've already worked on (even slightly) big volumes of data on Excel, you've probably experienced files freezing and/or circular references errors popping out of nowhere. These problems only get worse with big data. You'll be constantly dealing with a lot more (and typically a lot bigger) files, with even more data preparation steps when you make the switch.



When handling and treating big amounts of data, it's easier to think in terms of workflow. This means explicitly and clearly visualizing the pipeline your data goes through. You need to track and automate your data preparation steps to prevent you and your team from getting confused by the data analysis process and the connections between tables.

This will both save you time and help you make sure the information remains accessible and coherent. This will allow you to trace all the modifications your data goes through. And it will make reproducing various steps easy.

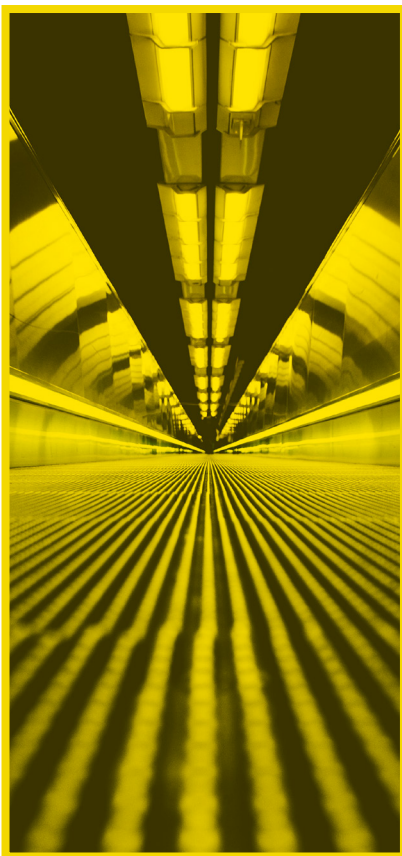


Key Takeaway

It's important to think collaboration as well when you're working with big data! With so much data to work on, you probably won't be going at it alone. Think about your current and future colleagues and set-up readable and documented workflows.



* Looking for Known Indicators vs. Discovering Unknown Trends: Unchained Possibilities



When working on small data, you generally have a specific goal in mind. You know exactly what information you have (transaction data, client information files, or product categories), so it's easy to define indicators to monitor over time or correlation you want to analyse.

Big data, by definition, is big. You store so much data from so many different sources that the range of information you have is much broader. This means that:

- 1/ you won't even realise everything that's being stored, and
- 2/ you don't know what part of what is being stored is actually of any value!

In big data, the crucial information can be in small signals hidden inside the data that you can't really identify by hand or even explain.

So beginning a project or an analysis up with a clear idea of the exact trend to look for in your data can turn out to be deceptive. Moreover, you could miss out on something really interesting. You have to be prepared to find more extra questions than answers, and work in an agile way to keep digging deeper.



Key Takeaway

That being said, this does not mean that you should dig in and explore your data with no idea of what you're looking for at all. It does mean that your objective and methods will be broader: you'll go in with a question, and try to find the answer.



Structured Data vs. Not Always So Structured: Begin the Treasure Hunt



The key to doing big data right is to use broad sources of information to enrich your data analyses. While small data often means siloed departments working with the data they collect, using big data tools allows you to break down the barriers between departments, and even bring in data generated outside the company.

This makes for a pretty awesome bunch of data. Consequently, it will also impact what your data looks like. Indeed, to be sure to find the needle in the data-stack, data scientists oppose important pre-processing of the data to be sure to lose as little of it as possible, so it comes out pretty dirty!

When working on exports in Excel for example, the data is generally pre-processed and cleaned, and comes out in neat fields in a record or file: that's what you call structured data. However when working with big volumes of data, your data will be much messier - with lots of formats, logs, timezones, and what's called unstructured data (text, video, image, metadata etc.)



Key Takeaway

You'll have to manipulate the data to extract info from it, and get it in a format that can be processed. Prepare for long nights spent cleaning text and dates to extract features (luckily there are tools to make it easier!) Don't get scared about how bad raw data can sometimes look, you'll get used to the treasure hunt - and you'll probably start enjoying it too.



* Ordering VS Data Visualization: Getting to Know Your Data

We're going to start exploring functions that shape the excel mindset now, and find the big data equivalent.

Let's start from, well, the beginning. When you start exploring data, one of the first things you use to explore your data is sorting. You'll order your data and get a quick view of what your data's like, fill in empty values, or delete them, etc. If you want to know how many of a product were sold today, you probably don't even have to look far, just sort, get the latest date at the top, and scroll a bit to get a general idea.

However, as soon as you begin working with big volumes of data, getting an idea of your transaction volumes can no longer be done just by scrolling the table you quickly ordered. Actually, you'll eventually have so much info that even stats and pivot tables won't be enough to see what's happening. That's when data visualizations become extremely handy.

When working on big data, making charts and graphs is not only the best way to visualize your data, it's the only one! "Data viz" (like the cool kidz call it) allows you to quickly and easily identify (and easily share) trends and patterns you would probably not have spotted just by looking at your dataset. It provides a quick answer to complex questions, regardless of the volume of your data.



Key Takeaway

Data viz is also a great way to communicate with people about what's going on with the data. That's why it's so hype these days, with tools like Tableau helping everyone make their own cool charts, graphs, maps, you name it. You can use them to get the data to say pretty much what you want, but also, if you're not careful, what you wanted to hide. So design your graphs carefully!



Filtering vs. Querying

Diving Straight In To Get Your Answers



Going from small to big data will also change the way you get the answers to the questions you're asking yourself, or the way you select the information you're actually interested in and want to work with.

In Excel, Google Sheets, and co, you're probably used to using filters or COUNT IFs on your data to get the information you need, all the while keeping your whole dataset at hand.

When working with big data you have to keep in mind that you'll always be working on all your data. Then, in all of that data, you'll select what you want to show or make calculations on or just ask your question. The way the SQL programming language works is actually quite representative of that mindset shift (and it's very accessible). You'll write a query, or a question - something like this: "select my clients who've been on my site in the last 30 days" or just ask "count how many of my clients have been on my site in the last 30 days." This query will be applied to the entirety of your data and return the data you asked for, or the answer you were looking for.



Key Takeaway

Spoiler alert! In the end, even if you're working with a little SQL you'll probably end up going back to Excel. But with languages such as SQL, you'll be able to export tables with your indicators all prepared, that you'll only have to copy paste to share with everyone or build graphs with!



* Pivot Tables vs Data Aggregations: Bringing It Down To The Essential Faster

Another essential feature of data mining in Excel is pivot tables. They're the best way to get statistics and results faster, or to dive into deeper data analysis.

The big data analytics equivalent of pivot tables is aggregations, or groupings. They allow you to separate data into groups, which can be aggregated independently of one another.

Groupings help you reduce the size of data by aggregating groups of observations to create summaries of the original data.

They can be used very similarly to pivot tables: they're both about detecting trends, making comparisons, and revealing insights you would not have noticed if you were looking at isolated data elements. And just like pivot tables you can create groupings with the minimum value, average, sum, number of distinct values, and lots of much more elaborate stats than what you'd get in excel.



Key Takeaway

Let's look at a concrete example. You've been asked to look at your company's website traffic. You're going to be working with logs of each visit for each page of your site. The first step to take is to group your rows by visitor ids for instance, to get a clear view of how many pages each visitor saw, when the last time he visited was, how many products he bought, etc.



VLookUp VS Joins

Finally Combine All Your Data Sources Stably

VLookUps and Index Match are the most important function for data analysis in Excel. They're fundamental to link together files and retrieve specific information from them- so you don't have to copy paste it everywhere any time there's an update. But if you have been using them for a while, you'll know how bad things get when they're everywhere.



Outside of excel- and especially in the basic database query language SQL- you'll have to get familiar with joins. People use joins to combine one or more datasets to create a table containing all the information you need from them.

These connections are very powerful and resilient because the SQL system was actually built to make these connections. How's that? Well, the point is to only have information stored in one place (or table) to save storage space. These tables are linked together with keys that serve as identifiers to create relationships among different database tables or views. To make sure that data is accessible, these joins have to be stable, making them much more powerful than Vlookups.



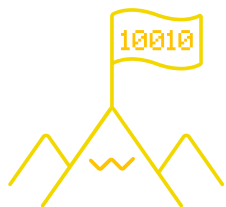
Key Takeaway

Joins can be used for a lot more than vlookups actually. There are several different types that allow you to select which is your main dataset, and if you want to keep only matching lines or not. You can also select a whole list of information you want to add easily, not just one.



Conclusion*

Big Data Analytics Without the Data Scientists and Engineers



The point of the matter is: it's time for analysts everywhere to take back the data. The technology and the tools today make it possible for anyone to take their analytics game to the next level by using the power of Big Data.



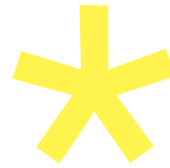
You no longer need to rely on data engineers to give you the data for every query you have. No need to wait for those massive amounts of data to be pre-processed and go through complex data cleaning pipelines. You can perform your analyses iteratively and in an agile manner by having the data available to answer the questions that will come with your answers.



The tools are available to ask your questions without requiring the intervention of a data scientist to translate into big data languages. Code free platforms allow you to be efficient, but also to gain new skills.



Everyone in a company who needs to have access to data to make better business decisions should have it. And without bringing in data scientists and engineers who could make better use of their time by creating predictive services and applications.



What Next?

This is it! Your first step to bring big data analytics to your job is done. So what now? Well, all the work is left to be done, but the fun part as well! The next step is to start digging into your data. This means going to your boss or your tech team to get access to the database and start exploring.

It means looking into tools to work efficiently. One option is to look into learning a new language to work with data. SQL is the easiest, and a basic. You can then move on to Python or R. There are lots of tutorials online, so I'm sure you'll find your way. Another option is to start looking into platforms with clickable interfaces to start finding your answers.





DATA SCIENCE STUDIO

- ✓ Work on all your data not just static extracts or siloed databases
- ✓ Perform data cleaning and exploratory machine learning without the help of a data scientist
- ✓ Learn new skills while still being productive
- ✓ Iterate on what you find and ask new questions
- ✓ Get fast answers on all your data sources
- ✓ Don't rely on data engineers to get you data
- ✓ Connect to data from over 25 storages systems (including Hadoop, Spark, NoSQL, Enterprise SQL..)
- ✓ Enrich with open source and social data thanks to plugins

Go to dataiku.com to download our Community version, free and yours to keep.

