7 Super Simple Steps From Idea To Successful Data Science Project

Ever had this great idea for a data science project or business? In the end you did not do it because you did not know how to make it a success? Today I am going to show you how to do it.

To keep this article as real as possible let's talk about a distinct use case.
The theory is that you can use twitter data to identify current big data trends.
The goal is to sell this service to customers for a topic of their choice. We are going to use AWS services to realise all of this.
I am not advertising for AWS, I just have more experience with it. I am sure Google Cloud or Azure have the same features, only named differently.

Step 1 - Get the data

Before you do anything else, you are going to need to prove your theory. To do that, you have to collect some data.
Twitter data can be accessed through public APIs. All you need to do is write a small program that can download tweets from certain users.
Use the programming language you are most familiar with. I for instance like Java a lot, so I use Java for these tasks.
When your software is working, download the tweets from all the influencers in the selected field. Influencers are people that have a large following and that tweet a lot of big data stuff.
For big data it is particularly easy because KDnuggets has a list of Twitter influencers: www.kdnuggets.com/2016/02/big-data-top-influencers-brands.html

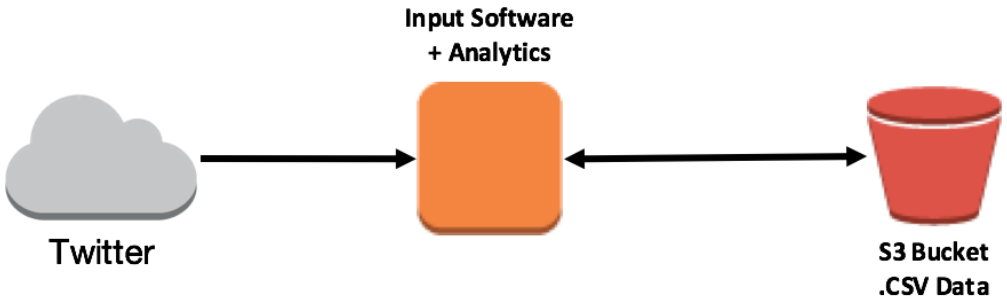
To do it in the cloud, you can spin up a simple AWS EC2 Linux instance (nano or micro), and run your software on in.

OK, but how should I store the data?

The best way to store the data is to use a simple .csv format. One line per tweet that includes the tweet's text and meta information.

The meta information you should include is person, time, replies, retweets and likes.

When you are finished, upload the file to S3.



How much data should I extract?

My advice is always to get as much data as possible in a reasonable time. Let your program run a few days.

Twitter has some strict API rules for how much data you can query in a certain time. You have to throttle your software to not go all out. It will get timed out otherwise.

Anyway, a few months of tweet history should be enough. How much data to get is not an exact science, follow your gut.

One more thing: Only collect as much data as the machine you are using for analytics can handle.

Step 2 - Select the right tools for the analytics

After getting the data you need to select the proper tool to analyze it. Write down a list of analytics features you think you need and compare available tools.

Cheap and quick is the way to go.

You could either go user friendly with graphical tools like Orange, Rapid Miner or Knime.

If they don't really fit then go the write the analysis yourself. Python and R are amazing languages for data science.

But I love using Matlab, can I use it?

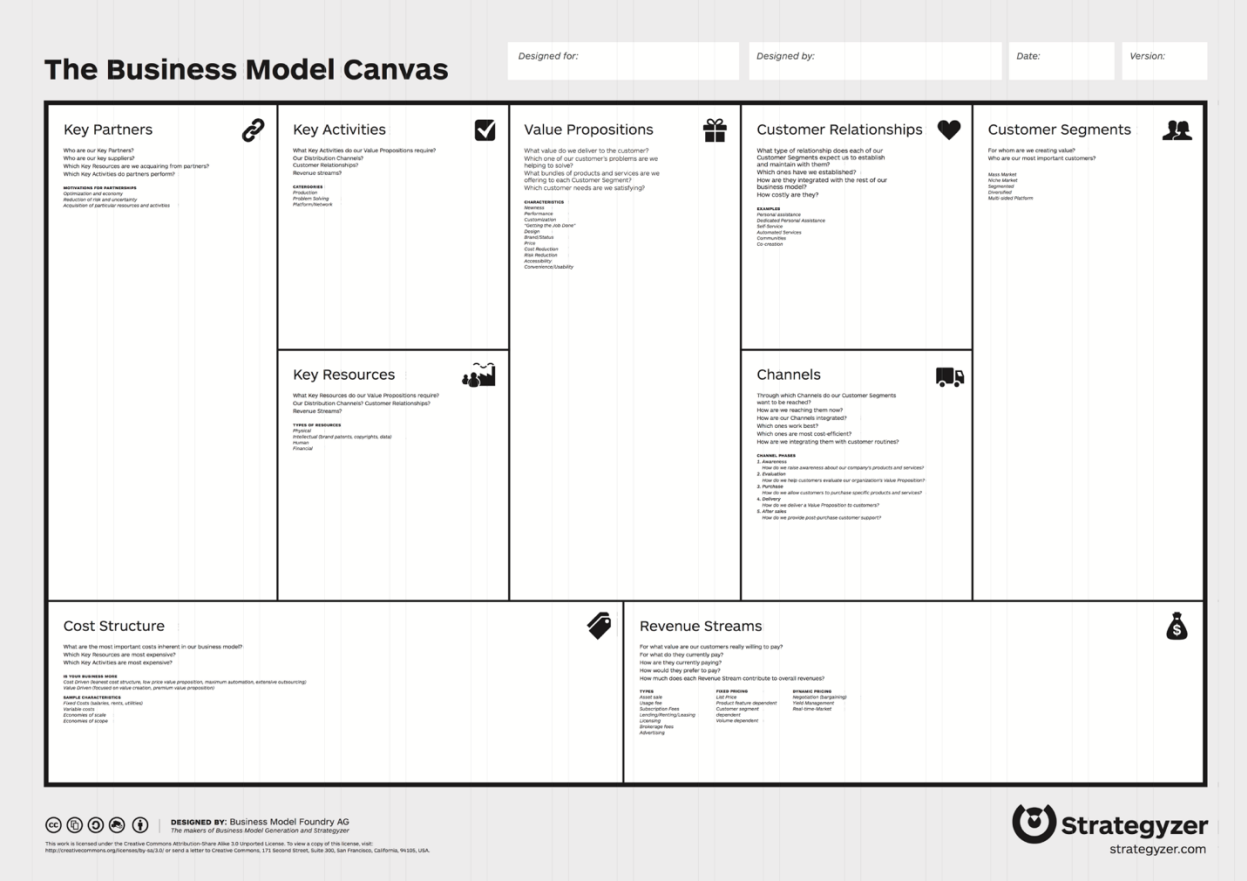
Of course, if Matlab has the features you need - use it! Use what ever fits your needs best.
You can use the previous EC2 instance to do the analytics on. Or stop the old one and spin up a new one if you need to change the operating system.

Step 3 - Prove your theory with science

You have the data and the tools in place. This means you are ready.
It's time to work the data and prove your theory!
But how?
Start by pinpointing trends you already know are in the data. A simple way would be to search google for significant events that have been reported much.
Try to create an analytics process that finds these trends.
How do I know when the theory is proved and it's time to move on?
If analytics can find the trends you specified, then you are on the right track. Look for instances where analytics finds new trends.
Confirm these trends, for instance by searching the internet. The results are not going to be reliable 100% of the time.
Before continuing you need to decide how much falsely reported trends (the error rate) you want to tolerate.
By the way 0% error rate is absolutely unrealistic.

Step 4 - Figure out your business model

After you got the science right, you should take a step back. Before continuing, you need to figure out your business model.
Ask yourself: What is it that you do, what resources do you need, and what value do you provide to the customer?
Who are your customers and how are you going to sell your product to them?
For what values are customers going to pay?
A nice way to do this is the business model canvas. It's simple and cheap, you can basically create it on a sheet of paper.
When you are finished, continue with building a minimum viable product (MVP).



ที่มา : https://en.wikipedia.org/wiki/Business_Model_Canvas

Infrastructure

- **Key Activities**: The most important activities in executing a company's value proposition. An example for Bic, the pen manufacturer, would be creating an efficient supply chain to drive down costs.
- **Key Resources**: The resources that are necessary to create value for the customer. They are considered an asset to a company, which are needed in order to sustain and support the business. These resources could be human, financial, physical and intellectual.
- **Partner Network**: In order to optimize operations and reduce risks of a business model, organizations usually cultivate buyer-supplier relationships so they can focus on their core activity. Complementary business alliances also can be considered through joint ventures, strategic alliances between competitors or non-competitors.

Offering

- **Value Propositions**: The collection of products and services a business offers to meet the needs of its customers. According to Osterwalder, (2004), a company's value proposition

is what distinguishes itself from its competitors. The value proposition provides value through various elements such as newness, performance, customization, "getting the job done", design, brand/ status, price, cost reduction, risk reduction, accessibility, and convenience/usability.

The value propositions may be:

- Quantitative – price and efficiency
- Qualitative – overall customer experience and outcome

Customers

- Customer Segments: To build an effective business model, a company must identify which customers it tries to serve. Various sets of customers can be segmented based on the different needs and attributes to ensure appropriate implementation of corporate strategy meets the characteristics of selected group of clients. The different types of customer segments include:
  - Mass Market: There is no specific segmentation for a company that follows the Mass Market element as the organization displays a wide view of potential clients. e.g. Car
  - Niche Market: Customer segmentation based on specialized needs and characteristics of its clients. e.g. Rolex
  - Segmented: A company applies additional segmentation within existing customer segment. In the segmented situation, the business may further distinguish its clients based on gender, age, and/or income.
  - Diversify: A business serves multiple customer segments with different needs and characteristics.
  - Multi-Sided Platform / Market: For a smooth day-to-day business operation, some companies will serve mutually dependent customer segment. A credit card company will provide services to credit card holders while simultaneously assisting merchants who accept those credit cards.
- Channels: A company can deliver its value proposition to its targeted customers through different channels. Effective channels will distribute a company's value proposition in ways that are fast, efficient and cost effective. An organization can reach its clients either through its own channels (store front), partner channels (major distributors), or a combination of both.

- Customer Relationships: To ensure the survival and success of any businesses, companies must identify the type of relationship they want to create with their customer segments. Various forms of customer relationships include:
  - Personal Assistance: Assistance in a form of employee-customer interaction. Such assistance is performed either during sales, after sales, and/or both.
  - Dedicated Personal Assistance: The most intimate and hands on personal assistance where a sales representative is assigned to handle all the needs and questions of a special set of clients.
  - Self Service: The type of relationship that translates from the indirect interaction between the company and the clients. Here, an organization provides the tools needed for the customers to serve themselves easily and effectively.
  - Automated Services: A system similar to self-service but more personalized as it has the ability to identify individual customers and his/her preferences. An example of this would be Amazon.com making book suggestion based on the characteristics of the previous book purchased.
  - Communities: Creating a community allows for a direct interaction among different clients and the company. The community platform produces a scenario where knowledge can be shared and problems are solved between different clients.
  - Co-creation: A personal relationship is created through the customer's direct input in the final outcome of the company's products/services.

Finances

- Cost Structure: This describes the most important monetary consequences while operating under different business models. A company's DOC.

  Classes of Business Structures:
  - Cost-Driven – This business model focuses on minimizing all costs and having no frills. e.g. Low cost airlines
  - Value-Driven – Less concerned with cost, this business model focuses on creating value for their products and services. e.g. Louis Vuitton, Rolex

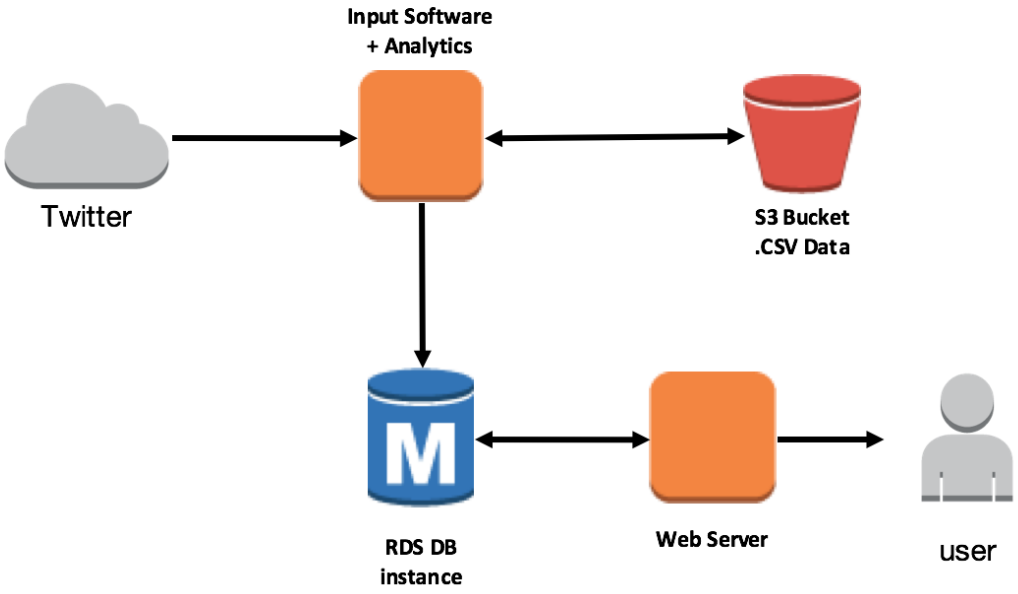
  Characteristics of Cost Structures:
  - Fixed Costs – Costs are unchanged across different applications. e.g. salary, rent
  - Variable Costs – These costs vary depending on the amount of production of goods or services. e.g. music festivals

- Economies of Scale – Costs go down as the amount of good are ordered or produced.
- Economies of Scope – Costs go down due to incorporating other businesses which have a direct relation to the original product.

- Revenue Streams: The way a company makes income from each customer segment. Several ways to generate a revenue stream:
  - Asset Sale – (the most common type) Selling ownership rights to a physical good. e.g. retail corporations
  - Usage Fee – Money generated from the use of a particular service e.g. UPS
  - Subscription Fees – Revenue generated by selling a continuous service. e.g. Netflix
  - Lending/Leasing/Renting – Giving exclusive right to an asset for a particular period of time. e.g. Leasing a Car
  - Licensing – Revenue generated from charging for the use of a protected intellectual property.
  - Brokerage Fees – Revenue generated from an intermediate service between 2 parties. e.g. Broker selling a house for commission
  - Advertising – Revenue generated from charging fees for product advertising.

Step 5 - Build a minimum viable product

After proving your theory, it is time to start building a first version called minimum viable product (MVP). The goal of a MVP is to build a solution that only delivers the core functionality.
Don't go for fancy solutions. Focus on the main functions you need to realize.
Stick to what you know and what will work in the beginning and expand your system later. It can be something really simple like a RDS database instance and an EC2 with Tomcat to deliver content.
The system could look something like this:



Basically, this can be the first version that you offer to customers. It has all the core features: extracting data from twitter, analyzing it and display the results to the customers.

Step 6 - Automate and Measure everything

A MVP usually does not only lack features, it needs automation.
Automate as much as possible. You need to be able to concentrate on the further development and not on system operation.
Automate how to upload data to S3, stop starting the analytics by hand and write an automation script.
Start the analysis automatically and no longer by hand.
Connect the download script to the RDS database to dynamically read the list of influencers. This allows you to automatically include new influencers on customer demand.
Automate everything, create API's to ingest and store data automatically.
Then there is logging.
You need to know what you should develop next. Not only in terms of new features, it's also about fixing problems with the platform and making it faster.
Set up a system for logging and monitoring. Try to measure as much as possible.
You can log server statistics like cpu, ram, network with tools like for instance Nagios. Nagios includes a user interface for these statistics.

Log statistics for the download from twitter or upload to S3.
Log how long the Analytics process is taking and other statistics.
Log what users are doing. A simple way is to write a line in the log every time a user is using a specific function of the user interface.

Step 7 - Re-iterate

So, now your MVP is running and you automated almost everything. You have a comprehensive monitoring in place.
The system is running for some time now, you know exactly how every aspect is behaving:
You know how fast you can ingest. You know the performance of the storage and the analytics.
You have a clear indication what customers are doing.
Because you implemented extensive logging, all the weaknesses in your design are visible.
What to do next?
It's time to further enhance your system. Get rid of the current weaknesses and add further functions to the system.
Getting rid of weaknesses you will optimize the overall performance and stability of the system.
New features will add further value for your customers.
Implementing new features will also allow you to offer new services or products.

Some Final Words

Is it really that simple to turn an idea into a successful venture? Just by:
Getting the data
Selecting the right tools for analytics
Proving your theory with science
Figuring out your business model
Building a minimum viable product
Automating and measuring everything
Re-iterating

Following these 7 steps will bring order into the chaos of building a product. They will help you to set your priorities and get the most out of your time.

However, to be honest: There is no success guarantee, despite people telling you that there is. There just isn't.

Maybe you misjudged your customers needs and you need to pivot by changing your customer value and how you deliver. If so, these seven steps will help you find your way.

What do you think? Did I miss a crucial piece of the puzzle? Have you already done the seven steps without knowing about them?

Bio: Andreas Kretz is a Data Science & Big Data Adventurer from Germany.

https: / / www. kdnuggets. com/ 2017/ 11/ 7- super- simple- steps- idea- successful- data- science-project.html