5.2.0 Failure

Hi, today we're going to be talking about failure in big data systems and big data analysis. There's a very famous quote from a German battle commander, who said no battle plan survives first contact with the enemy and he's absolutely right.

More Big Data project failed and succeed. And there's a number of reasons for this and this is a hotly debated topic. And if you go and talk to people that are involved in specific Big Data rollouts, or Big Data Analysis projects, they will all have their own theories on why the project that they were involved with, either failed or wasn't as successful as they'd hoped it would be. So, we're going to have a little look at some of the reasons, some of the most common reasons, that people give for why these projects fail.

So, people, now people has a number of connotations to it, this could be the wrong people involved. Most Big Data projects are led by data scientists and data analysts. This leads to mismanagement in the project, it leads to a lack of understanding about how the project itself may fit in with the business needs of the day, because big data analysts and data analysts in general, are trained to look at the data in a scientific way. They're not trained to apply bias to it intentionally. So people mismanagement, just having the wrong skills, just because somebody is capable of building a system that analyses small amounts of web traffic, and does basic analytics on location doesn't mean that they can automatically build you an entire health care system for the NHS, or for a health service or hospital, those two things are worlds apart. So just because somebody who has experienced in one of the areas that might be required doesn't necessarily mean that they can do everything in it. So, people are a big one.

Time, a lot of these projects run out of time, they tried to solve too many problems, they have what's called scope creep, they have problems where they're just they continue to build, and they never really launch or they create, they spend all of their time creating, this brilliant analytical system, and they don't have the data to put into it.

Poor planning, you know, poor planning really comes down a lot of the things we're going to cover here come down to poor planning and can be addressed through planning. But poor planning is definitely one of the big things people will look at typically businesses, especially those kinds of medium sized companies, they will look at the new buzzword of big data and data analytics. And they will say, Oh, we should be doing that and you say, Okay, what should you be doing with that, and they said, we should be doing that, we should do some big data analytics, we should do data analytics more. And they'll run off and they'll employ people to do it, they'll give it a bit of budget, though, you know, put some time on it. But they don't define and plan exactly what outcomes they want. They'll go into it more with the idea of well, we'll do it and we'll see what we find out what use it’s going to be at the other end.

Legal issues are another big challenge here and a lot of projects fall down, when they start to collect things like sensitive personal data. They don't do anonymization correctly. They don't do pseudonymise correctly. There's a number of legal constraints about where you

can store somebody's data, where you can process somebody's data, how you can move data from one place to another and these legal issues will be covered specifically, in later sessions, but they have a big bearing here.

Security is another thing that he's thrown out time and time again, we built this amazing system, but we needed it to be fast and we had deadlines. So, we've built it, and now it's been hacked, somebody has gained access to it. And now we've got to shut it down.

Bad data or insufficient data coming into the system, I've built an amazing thing. It's like going out and building a Ferrari, but then forgetting to put petrol in it, or putting diesel in your brand-new petrol Ferrari, this is just never going to work. You have to understand where your data is coming from and you have to have a plan for how you're going to deal with it, when it arrives, and how you're going to continue to feed data into it if that's what it requires.

And also funding. Funding runs out for these projects, especially if they're poorly planned, especially if they exceed time expectations, or they encounter legal issues, or security problems, these will suck up funds from a project. And all of a sudden the funding that was given and that was earmarked to go to the big data analysis project is suddenly gone and the whole project collapses.

Another problem faced by new big data systems, and really any data analysis systems, it's not solely down to big data at this stage is that it's not really a thing yet. People don't really have a common understanding of what we mean by big data analytics.

So, a simple example of this is, what does the ‘big’ mean in big data analytics?

Do we mean 1000 data points? Do we mean 1000 data sources? Do we mean 10 data sources with 1000 data points each. There isn't really a defined line in the sand that says, okay, below this, we're doing analytics. Above this, we're doing big analytics, that line in the sand moves and it changes depending on who you talk to and it changes most of the time depending on what you're analysing.

So to give you an idea of what we mean by not really a thing yet, when you come up with a new field, when you come up with a new application for an existing field, which is what a lot of the big data analytic stuff is. If you want to apply this idea, you want to get this idea you want to get it recognised, and you want this idea to grow up and become a recognised thing a recognised industry recognised defined model, whatever it happens to be, you start with the idea, you have the idea.

The idea then creates, or has grown around it, some recognised practices. People start doing things in a certain way, people start to be able to share what they've done, people start to be able to replicate what was done and we end up with a recognised way of doing things and some recognised practices for dealing with it.

Once we've got those recognised practices, we then kind of move on and we say, Okay, now that everybody sort of understands what this is, we've got some lines in the sand. Let's write

those down. Let's define those as rules of what is and what is not big data analytics. And what is something that you should do and isn't something that you should do? How is it done? It then moves on and becomes a measurable standard.

So now that we've got our rules, we can then measure people against those rules at a certain point and say, Okay, are you doing it well or are you doing it badly? Are you doing it in line with what the current industry standard is? Are you doing it in line with how an expert, a recognised expert, has said it should be done? So can we measure what's being done against some sort of rule, and some sort of standard and some sort of expert opinion, to see whether somebody is doing it well or badly once we reach this level, and once our idea is taken this journey through all of these steps, we end up at this point where it's a real thing, and it's a real thing, because people that don't work in it, people that have passed by it, people that only have a passing understanding of it, can point at it and say that's what it is, they can look at it and they can say it's a real thing and I know it's a real thing because it's defined as this. I can go and look it up on the internet, I can go and look it up in the dictionary and there is a definition next to it. That makes sense in all cases.

We haven't really reached this with big data analytics yet, we're kind of still in this defining the rules stage somewhere between recognised practices and defining the rules. Because it's quite a young art form, it's a you know, it's certainly a young science and it's going to grow up, it seems to be heading that way. We've seen other things that have emerged and then kind of fallen by the wayside. But big data analytics seems to be growing, as we've seen with the challenges in previous sessions, as we've seen with the problems that the world is facing at the moment, we know we need this, but it's still trying to find its feet and it means that a lot of the projects that started out as just an idea, a lot of the projects that people say, Oh, yeah, let's use big data analytics, let's do some of that, they get as far as trying to implement it, and never really implement anything, never really get it right. And a lot of that also boils down to this fact that there isn't a measurable standard and it's not necessarily recognised universally, as the same thing, or as the same understanding that everybody else has of what it is.

Fail to plan, plan to fail. This is a very famous quote from Benjamin Franklin. Although I did a little bit of research into this and apparently, he never actually, categorically said this, there's no evidence that he ever said those words. But it was attributed to him by a newspaper in 1970. And everybody has believed that ever since that it was him. So fail to plan, plan to fail, it still makes sense.

So, because Big Data is an emerging science, there's no agreed formula for if you do these things, you will be successful in it. If you go through these motions, it will always work. There is no A plus B equals C. It has to be defined for every single implementation and every single project specifically. So how can we do that? Here's some rules of thumb for to help guide you, as you're looking at and planning these implementations.

The first one is have a clear objective, formulate your objective around a very specific question, the more specific your objective and your question can be, the more likely you are to follow it through, treat that as your true north and not get lost along the way.

Understand, going into a project, what resources are available to you for that project and for how long you will have those resources. Crunching large data is expensive. Doing full analysis is not something typically that a single individual is going to sit down and do. This normally takes a team, you're going to have to build things, you're going to have to have the relevant skill sets, you might have to have computational time that's paid for, you're going to want to be able to back up the data that you're using, you're going to want to be able to create persistence, all of this costs money, all of this costs time. All of this is a drain on resources. So, what resources do you have available? And how long are you likely to have those resources for?

At this point, you are able to create a plan for gaining a full understanding how you're going to get to your objective. So, using these resources for this amount of time, again, to achieve this objective, by answering this specific question.

Now that you have a plan, be prepared to abandon your plan. Okay, be prepared to adapt your plan. As we said before, no battle plan ever survives first contact with the enemy. Be prepared to adapt, have contingencies in place for predictable failures. If you have a plan that relies on you receiving x amount of data into your data set per hour per day. What happens if that goes down for two days? What's your contingency? Can the system still function? does everything fall apart? These are things that are predictable.

What happens to you if the system crashes? What happens if the system hits some sort of a parsing error? What happens if there's a bug in the code and everything needs to be shut down? Okay, what happens if people go off sick? What happens if somebody decides to quit the project and leave and move on? Do you have contingencies in place for these predictable failures? It's required if you want the project to succeed or not wither on the vine.

Once you've got your project up and running, and you're starting to produce output, you're starting to create results, you're starting to draw conclusions. Make sure that you gain external validation on this. Because there's a lot of results coming out at the moment, especially in things like the medical field, political fields, banking investments, there's a lot of results being published at the moment, but very, very few of them are peer reviewed. even fewer of them can be replicated. So, make sure that if you have a result that you want the world to see. Make sure that it’s externally validated in some way.

And finally, here, communicate clearly throughout development and execution of the process. A lot of times, we'll see brilliant communication upfront when the design phase is happening and when all the ideas are flowing around the table, and everybody's excited about this new thing that's going to happen. And yet this time next year, we're all going to be millionaires, because we're going to create these great analytics, and everybody's going to need it or, you know, we're going to solve world hunger. But all of a sudden, somewhere in the development, challenges start to crop up, things start to go wrong and that communication that was so good at the beginning, slowly starts to fall apart and by the time your project gets to that execution and implementation phase, there's very, very little communication.

This results in the people funding the project starting to really review that funding and saying, “Do we know what's happening with all that money we're pouring into that over there? Should we be pouring money into it? Are we seeing value back from it? Do we know what's happening to the people that are involved in it, demonstrate to us that they that they should have our faith and our trust to continue with this process?”. So, make sure that the communication is clear, throughout the entire process throughout the entire project, not just at certain stages.

So, what can we learn from failure? What can we do with failure?

Another famous quote here, when asked about the light bulb, and the and, and why he kept trying to create the light bulb, and after failing 2000 times, the response was simple. I never failed to make a light bulb, I found 2000 ways not to make a light bulb. This is one of Thomas Edison's most famous quotes, because it was an inventor and entrepreneur, a scientist embracing failure and it demonstrated that failure is often one of the most important steps on a journey to succeeding and we, as data scientists should embrace these failures.

Big Data, as we've said, is young as a field young as a scientific field. It's essential therefore that we learn the lessons that we can, when failure happens, because taking the time to really dissect what happened. And why means that we can not only be better prepared ourselves, but we can share that information amongst other data scientists, and the field can improve at an exponential rate.

We should also when we're planning our projects, and when we're creating new ideas and new projects, and even if we've got a successful project running, we should definitely look to previous projects, things that have come before us for the lessons that we can learn. And when I say look for lessons, I mean, in the good and the bad. So, whether it was things like the NHS, the UK NHS informatics service, trying to go digital, which is very, very famously not gone. Well. It was a £17 billion project. By the time the government pulled the plug, and it was never implemented, and it was never finished. You know, you've got other examples like Lego who trusted big data analytics and change their entire business strategy around it and nearly went bust. You know, but we can also look at the good. You've got people like Kaggle, who are now using crowdsourcing to handle big data analytics. You know, using that big resource pool to deal with big data in an innovative way that allows the best to come out of it. So, look at good and bad projects previously.

Document everything. It may seem insignificant time that you're doing it, it may seem like you need to stop because something's going wrong and you haven't got time to write down what's going on, you haven't got time to document what's happening, you haven't got time to make sure that all of those communications are filed away properly. But trust me, it will prove its weight in gold, when you have to go back and look at or dissect what went wrong, and what could have been done to fix it, or what could have been done to avoid issues.

Finally, backup everything. Okay, systems fail, hardware fails, software fails, if you are putting things up in the cloud and using cloud processing, you have no control over those systems. If, you know, something like a ransomware, were to make its way into your

network. If hackers get their way in there and start deleting things, whatever is going on, you know, bad actors, insider threats, if anything was to happen to your project part way through and you had to start again, ask yourself whether that would be possible.

Backup everything as you go document everything as you go, keep change logs, keep decision logs, make sure that you understand what's happened, and how to replicate it. When you want to go for external validation. You want to be able to hand over how you did things so that they can replicate what you did, and make sure that they get the same results.

So, this concludes our session on failure in big data. Next, we will be visiting reanalysis and talking about its importance