How to determine ‘Best Practice’ in a Big Data Project

How to determine “Best Practice” in a Big Data Project

Big Data, and IT in general, is a huge field. There are often many ways to achieve the same outcome.

I often hear the question: *Which of these ways is “best practice”?*

This can be a hard question to answer as all projects and organisations are different. There is rarely one best practice that applies in all situations. Here are some strategies for ensuring that good practices are implemented on your projects.

Eliminate bad practice options

The first thing to do is eliminate the approaches that are clearly bad practice.

eg. Don’t rely on people remembering to turn off compute resources when no-one is using them. People will inevitably forget from time to time. You will pay a fortune and get no value from it. Instead, work out when compute resources are needed and set up automated processes to shut them down when not in use.

Some ways to identify bad practices are:

- Components that are more expensive than the value they add e.g. using oversized cloud resources
- Components that will limit the solution in some way e.g. reports that will only work on a small dataset
- Adding new components that are not scalable to reasonably foreseeable workloads e.g. using a SQL database when you know that within a year the workload will grow to billions of rows
- Components that will mean a solution is not flexible / extendable e.g. it’s not simple to add new data items to a data model
- Components that may have security issues e.g. using an individual’s user account to run production resources, instead of using a resource specific security credential
- Elements that will mean the solution becomes more challenging to support e.g. unnecessary complexity and poor documentation

Review good practice options

Once you have eliminated the bad options, you will frequently still be left with several options which could be considered a good practice approach.

This is where it gets tricky. You need to consider not just Big Data approaches in general, but the specific problem you are trying to solve.

Questions to ask:

- *Is this a "greenfields" application or is there an existing IT Infrastructure that you are adding to?*
  In the case of existing IT infrastructure, do you already have technologies in use that could be used for your project? Does a new technology add enough value to justify the learning curve? (it might).
- *Is there an existing IT team, and if so, what technologies are they familiar with?*
  For example, if the team has mostly been working in SQL Server, using GUI based interface tools, then using a solution with a mainly command line interface that only works with python is going to be a very steep learning curve. A different solution with a GUI and the option to run SQL based queries is going to make adoption by the existing team much faster and smoother.
- *Do you have restrictions such as legislation (eg. relating to Data Sovereignty)?*
- If all your data must remain in Data Centres within your country's borders, then a solution that pushes data offshore can be ruled out immediately.
- *Does the organisation have a strategic roadmap for IT and Big Data?* If so, how can your project move the company in the direction of the roadmap?
- *What is the budget for your project? How about your ongoing operations budget?*
- *Are there any vendor best practices that are relevant?* Microsoft, AWS and Databricks all offer several reference architectures and best practices that can be applied to many projects.

All of these considerations, and many project specific ones, need to be weighed up. This can lead to "Decision Paralysis".

Making a list of your requirements, constraints and existing resources and assessing your options against them in a structured analysis can help with this. It is important to document your reasoning.

Most importantly: **Don't let perfect be the enemy of good**. Modern Big Data toolsets, such as Databricks, Microsoft Fabric and AWS Lake Formation, are designed to be modular and flexible, so you can adjust details of your approach without throwing away the whole existing setup. Once you have identified a viable approach, start building (eg. a proof of concept). Be ready to reassess direction regularly and course correct if you need to.

Fujitsu Data & AI are specialists in setting up Big Data Solutions across a variety of platforms and technologies. Please contact one of our Data & AI specialists by emailing us or call **03 9924 3000**, for a complementary consultation to see how we can help you work through the available options and pick the best approach for your situation.

Contact

Fujitsu Data & AI
+61 3 9924 3000