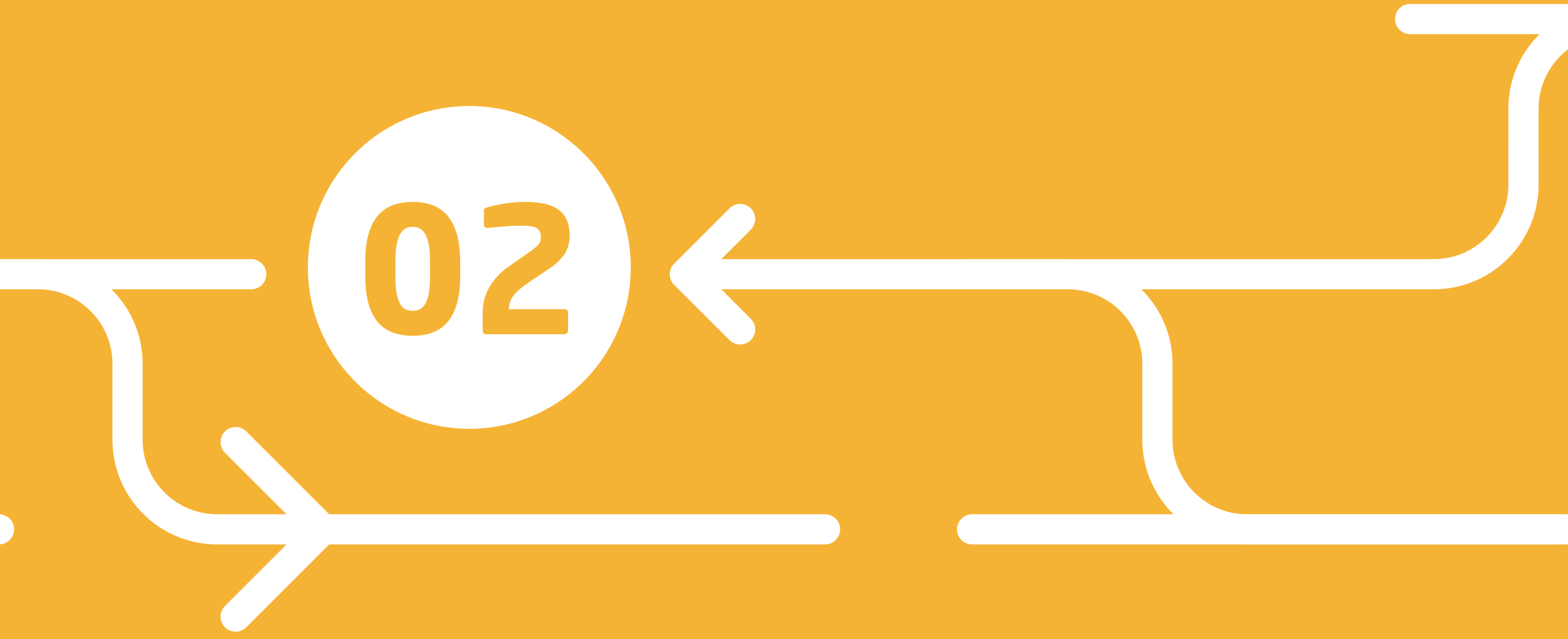
مركز محمد بن راشد للابتكار الحكومي
MOHAMMED BIN RASHID CENTRE FOR GOVERNMENT INNOVATION

ابتكر
IBTEKR

How to Gov Series

How To Approach Data Analytics Project Development



GOVERNMENT OF
UNITED ARAB EMIRATES

Info@pmo.gov.ae

TABLE OF CONTENTS

Chapter Four

04

How to Design and Implement a Data Analytics Project

180 pts
19

If a department, agency, or city is considering using analytics, there are a few ways to get started. While some organizations follow general standards of practice to provide a step-by-step guide on the key phases of project development, others follow a technical framework to identify the degree to which organizational and/or data resources will support a given project.

This Chapter highlights five key steps that you can replicate to develop you own analytics projects:

1. **Identify the problem**
2. **Assess data readiness**
3. **Scope the project**
4. **Pilot the project; and**
5. **Implement and scale the model.**

1. Identify the Problem

Identifying a critical problem that can be supported or alleviated by analytics is challenging, but it is an important first step in structuring a successful analytics project. While data may abound, matching an area of need with the right data resources within an organization is vital. Developing an analytics project typically places data scientists in an internal consultant role; by working with a department or agency to identify their key issues or problems, data scientists can support mission-critical needs.

To help narrow down a problem, it's helpful to understand five problem types where data analytics is particularly well suited to help:

1) Targets are difficult to find in a larger population

Example of a Data Analytics Product that could help with this problem: A graph of the population showing anomalies or outliers

2) Services do not categorize high priority case early

Example of a Data Analytics Product that could help with this problem: An algorithm that generates a prioritized list of cases

3) Resources are focused on reactive services

Example of a Data Analytics Product that could help with this problem: An algorithm that generates alerts to flag issues when a threshold has been reached/before damages have exacerbated.

4) Repeated decisions made without access to all relevant information

Example of a Data Analytics Product that could help with this problem: Data visualizations that help in accessing and understanding all relevant information

5) Assets are scheduled or deployed without input of latest service data

Example of a Data Analytics Product that could help with this problem: A map or heat map showing where cases occur

2. Assess Your Data Readiness

Determining data readiness is a key facet of analytics and a critical precondition to scoping any project. The success of an analytics project depends not only upon whether there is a need for data analytics, but also, and more importantly, on having the right personnel, data collection and storage practices, and stakeholder buy-in within and outside of the department or agency.

A helpful way to determine data readiness is by developing a Data Maturity Framework that consists of a questionnaire and scorecards to identify the technology, data, and organizational readiness within a department. The framework can include a questionnaire and survey to assess readiness and three scorecard matrices on:

1) Problem definition,

2) Data and technology readiness, and

3) Organizational readiness.

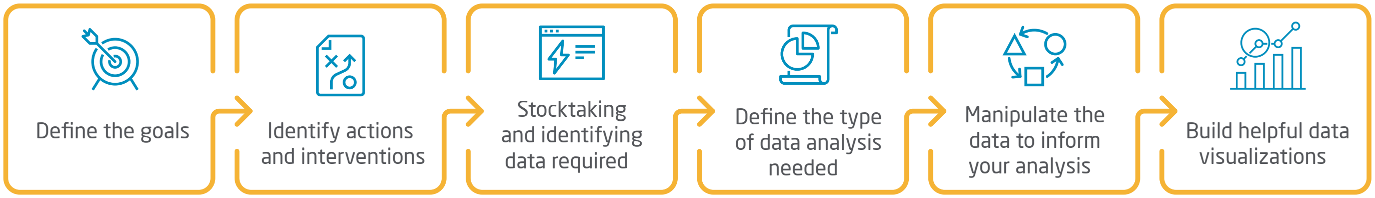
These scorecards can help organizations identify where they fall on a spectrum of four categories ranging from leading to lagging in terms of data readiness. Scorecard categories can include: how data is stored; what is collected; privacy and documentation practices; personnel; data use policy; and buy-in from staff, data collectors, leadership, intervener, and funder.

Assessing data maturity can also be approached from the macro level—for a mayor or chief data officer to assess the enterprise-wide maturity of municipal data, it is important to consider broad-scale questions such as how a government consumes data and how leadership uses data to make policy decisions.

3. Scope the Project

Once a department's data readiness is assessed, it is time to scope the project. There is no one "right" approach to project scoping. As always, the scoping process is fairly iterative and the scope gets refined multiple times both during the scoping process as well as during the project.

Below is an example of steps that can help you with scoping the problem:



Define the goals
Identify actions and interventions
Stocktaking and identifying data required
Define the type of data analysis needed
Manipulate the data to inform your analysis
Build helpful data visualizations

Step 1: Defining the Goals

Define the goal(s) of the project.

Step 2: Identifying actions and interventions

What actions/interventions do you have that this project will inform?

Step 3: Taking stock of existing and required data

What data do you have access to internally? What data do you need? What can you augment from external and/or public sources?

Step 4: Defining the type of analysis needed

What analysis needs to be done? Does it involve description, detection, prediction, or behavior change? How will the analysis be validated?

Step 5: Manipulating your data

How can you extract what you need from your data?

Step 6: Data Visualizations

How can you explore and present your data in a way that is easy to understand?

7. Ethical Considerations :

What are the privacy, transparency, discrimination/equity, and accountability issues around this project and how will you tackle them?

Step 1:
Define the Goal(s)

This is the most critical step in the scoping process. Most projects start with a very vague and abstract goal (say, improving education or healthcare), get a little more concrete (increase % of percentage of students who will graduate on time or decrease the number of children who get lead poisoning), and keep getting refined until the goal is both concrete and achieves the aims of the organization. This step is difficult because most organizations haven't explicitly defined analytical goals for many of the problems they're tackling. Sometimes, these goals exist but are locked implicitly in the minds of people within the organization. Other times, there are several goals that different parts of the organization are trying to optimize. The objective here is to take the outcome we're trying to achieve and turn it into a goal that is measurable and can be optimized.

Lets look at an example here from Higher Education in the US. One of the bigger challenges US High Schools are facing today is helping their students graduate (on time). Graduation rates in the US are ~65%. They're all interested in identifying students who are at risk of not graduating on time. When initially talking to most school districts, they start with a very narrow goal of predicting which kids are unlikely to graduate on time. The first step is to go back to the goal of increasing graduation rates and asking if there is a specific subset of at-risk students they want to identify. What if we could identify students who are only 5% likely to be at-risk versus students who are 95% likely to not graduate on time without extra support? If the goal is just to increase graduation rates, the first group is (probably) easier to intervene with and influence while the second group may be more challenging due to the resources they need. Is the goal to maximize the average/mean/median probability of graduating for a class/school or is the goal to focus on the kids most at risk and maximize the probability of graduation of the bottom 10% of the students? Or is the goal to create more equity and decrease the difference in the on-time graduation probability between the top quartile and the bottom quartile? All of these are reasonable goals but the schools have to understand, evaluate, and decide on which goals they care about. This conversation often makes them think harder about analytically defining what their organizational goals are as well as tradeoffs.

Considering tradeoffs while deciding on goals

As we start determining and often prioritizing goals, the conversation leads to tradeoffs. When dealing with students who may need extra support to graduate on time, what do you care more about? Finding every single student who may need that help (at the expense of targeting students who may not need the support and possibly being inefficient) or prioritizing efficiency and only focusing on students where you're extremely sure they'll need the extra support (and thus missing many students). Would you rather inspect more homes without finding lead hazards in them (inefficient) or would you rather miss homes with children who will end up getting lead poisoning? When dispatching and placing emergency response vehicles, do you want to make sure you can get to every possible emergency within 10 minutes or do you want to make sure that you can get to critical emergencies within 3 minutes and the non-critical within 20 minutes? What mistakes are you willing to make? That is a critical question a good scoping process brings up and answers based on the priorities of the organization.

In data analytics terms, would you rather have more false positives or more false negatives? Of course, this decision depends on the impact and cost of those errors, which is often hard (and sometimes uncomfortable) to quantify. There may not be an objectively correct answer but policymakers need to decide which policy goals they want to optimize, what resources they have, and which outcomes they want to prioritize. The data science work is then used to support and implement those policy goals. Data Analytics can help explore the impact of those goals and understand the implications better but it's ultimately a policy decision to decide on what goals to optimize.

Step 2:
What Actions/Interventions are you informing?

The work we do can typically only have impact if it's actionable. What actions can the organization take to achieve these goals? These actions often need to be fairly concrete: home inspections, enrolling a student in one of three after school programs, targeted emails for fundraising or advocacy, dispatching an emergency vehicle, or scheduling a waste pickup. A well-scoped project ideally has a set of actions that the organizations is taking that can be now be better informed using data science. If
the action/intervention a public health department is taking is lead hazard inspections, the data analytics work can help inform which homes to inspect. You don't have to limit this to making existing actions better. Often, we end up creating a new set of actions as well.

Generally, it's a good strategy to first focus on informing existing actions instead of starting with completely new actions that the organization isn't familiar with implementing. Enumerating the set of actions allows the project to be actionable. If the analysis that will be done later does not inform an action, then it usually (but not always) does not help the organization achieve their goals and should not be a priority.

Breaking Down Actions

Actions have a granularity, frequency, time horizon, channel, etc. For example, we would want to determine what channel(s) (Door knock, Phone call, Email, Twitter, Facebook, Snapchat, TV Ads) to use to target an individual? How often should they be targeted? Who should target them? You would also want to often come up with new actions and interventions.
Let's look at some additional examples of actions:

Often, an organization has one high level action (for example, lead inspection or home inspection or after school programs). In the scoping process, we can proceed in two ways:

1. We can just keep the scope to informing that one action. For example, which homes to inspect for hazards? Or which students should be enrolled in the after-school program?

2. We can also break the high level action down into smaller components: There may be multiple after school programs and each of them can be considered an action. For example, there may be 3 types of programs:

a. An online program that can be provided to 90% of the students

b. A short program that can be provided to 50% of the students

c. An intense program that can only be provided to 10% of the students.

Step 3:
What Data do you have and What Data do you need?

You'll notice that so far in the scoping process we haven't talked about data at all. This is intentional since we want these projects to be problem-centric and not data-centric. Yes, data is important and we all love data but starting with the data often leads to analysis that may not be actionable or relevant to the goals we want to achieve. Once we've determined the goals and actions, the next step is to find out what data sources exist inside (and outside) the organization that will be relevant to this problem and what data sources we need to solve this problem effectively.

For each data source, it's good practice to find out how it's stored, how often it's collected, what's its level of granularity, how far back does it go, is there a collection bias, how often does new data come in, and does it overwrite old fields or does it add new rows?

You first want to make a list of relevant data sources that are available inside the organization. This is an iterative process as well since most organization don't necessarily have a comprehensive list of data sources they store.

Sometimes, (if you're lucky) data may be in a central, integrated data warehouse but even then you may find individuals and/or departments who have additional data or different versions of the data warehouse.

Matching the Data to the Actions

This step also helps you figure out if your data matches the actions you need to inform. If the actions are individual level, then you most likely need data at an individual level. If the actions need to be decided on once a day, then you need your data to be updated every day. It's important to match the granularity, frequency, and time horizon of the actions to the granularity, frequency, and time horizon of the data you have.

External and/or Public Data

Once you've determined what data you need and what data exists inside the organization, you then want to figure out what external and/or public data you can get that fills the gaps. Each domain often has commonly used data sources that you want to know about. Open data portals (at federal, state, and local levels) also have data that can be used to augment your internal data. You also want to take a look at commercial data sources you can buy to augment your internal data.

Explore and Clean Your Data

Once you've gotten your data, it's time to get to work on it. Start digging to see what you've got and how you can link everything together to achieve your original goal. Start taking notes on your first analyses and ask questions to business people, the IT team, or other groups to understand what all your variables mean.

The next step (and by far the most dreaded one) is cleaning your data. You've probably noticed that some data is missing or incomplete. It's time to look at every one of your columns to make sure your data is homogeneous and clean. This is probably the longest, most annoying step of your data analytics project. It's going to be painful for a little bit, but as long as you keep focused on the final goal, you'll get through it.

Step 4:
What Analysis Needs to be Done?

The final step in the scoping process is to now determine the analysis that needs to be done to inform the actions using the data we have to achieve our goals.

The analysis can use methods and tools from different areas: computer science, machine learning, data science, statistics, and social sciences. One way to think about the analysis that can be done is to break it down into 4 types:

1. Description: primarily focused on understanding events and behaviors that have happened in the past.

2. Detection: Less focused on the past and more focused on ongoing events. Detection tasks often involve detecting events and anomalies that are currently happening.

3. Prediction: Focused on the future and predicting future behaviors and events.

4. Optimization

5. Behavior Change: Focused on causing change in behaviors of people, organizations, neighborhoods. Typically uses methods from causal inference and behavioral economics.

There are of course many more types of analysis but we'll keep the focus on these five in this manual.

The questions to answer in this step are:

1. What analysis needs to be done? Is this a descriptive analysis, a predictive model, or a detection or behavior change task? Often, the analysis involves several of the types of analysis we described above

2. How will the analysis be validated? What validation can be done using existing, historical data? What field trial needs to be designed to validate this in the field before it can be deployed?

3. How will the analysis inform the actions?

Step 5:
How can you manipulate the data for your analysis?

Now that you have clean data and you have figured out what type of analysis to run, it's time to manipulate it in order to get the most value out of it. You should start the data enrichment phase of the project by joining all your different sources and group logs to narrow your data down to the essential features. Another way of enriching data is by joining datasets – essentially, retrieving columns from one dataset or tab into a reference dataset. This is a key element of any analysis, but it can quickly become a nightmare when you have an abundance of sources.

When collecting, preparing, and manipulating your data, you need to be extra careful not to insert unintended bias or other undesirable patterns into it. Indeed, the data that is used in building machine learning models and AI algorithms is often a representation of the outside world, and thus can be deeply biased against certain groups and individuals. One of the things that make people fear data and AI the most is that the algorithm isn't able to recognize bias. As a result, when you train your model on biased data, it will interpret recurring bias as a decision to reproduce and not something to correct.

This is why an important part of the data manipulation process is making sure that the used datasets aren't reproducing or reinforcing any bias that could lead to biased, unjust, or unfair outputs. Accounting for the machine learning model's decision-making process and being able to interpret it is nowadays as important a quality for a data scientist, if not even more, as being able to build models in the first place.

Step 6:
How can you build helpful visualizations?

Now have a rich and organized dataset (or maybe several), so this is a good time to start exploring it by building graphs. When you're dealing with large volumes of data, visualization is the best way to explore and communicate your findings and is the next phase of your data analytics project.

The tricky part here is to be able to dig into your graphs at any time and answer any question someone would have about a given insight. That's when the data preparation comes in handy. If you are the person who did all the data collection and analyses work, you will know the data like the palm of your hand!
Graphs and charts are also another way to enrich your dataset and develop more interesting features. For example, by putting your data points on a map you could perhaps notice that specific geographic zones are more telling than specific countries or cities.
Ethical Considerations:

The ethical issues need to be considered throughout the scoping as well as the project execution process. Often the initial conversation around ethics will be around the values we want the solution we're building to have. This is not about data analytics or AI but about the social values we want to have. We need to make sure that this is not an afterthought or a burden, but rather a critical and continuous area of focus, and that involves all stakeholders, especially the people who are going to be impacted by this system.

Some of the issues that need to be discussed include:

- **Transparency**
  - Do the people who "own" the data know you're using it?
  - What actions are you taking on individuals based on this data?
  - Do the people you're "targeting" know why and if they're being "targeted"?
  - What recourse do they have?
  - Would it be an issue if they found out what you're doing?
  - Which stakeholders should know about which parts of the project?

- **Discrimination/Equity**
  - Which specific groups for whom you want to ensure equity of outcomes?
  - How would we you define equity? Over what period of time?
  - How do we detect inequities?

- How do we reduce inequities?
- How do we mitigate the equity impact of a biased system?

- **Accountability**
  - Who are the people responsible and accountable for all the things above?

- **Social License**
  - If the entire population of the country finds out about your project, will they be ok with it? Why or why not?

- **Privacy/Confidentiality and Security**
  - What are the privacy considerations (legal as well as ethical)?
  - How is the privacy of the individuals in the data being protected?
  - What about confidentiality?
  - What are the security considerations and protections? Who has access to which parts of the data? For what purposes? What is the security audit process?

- **Any other ethical considerations?**

4. Pilot the Project

Piloting an analytics project is “the stuff of innovation.” This is where the trial and error of testing a new project happens. No matter how well prepared an analytics team is, sometimes—whether the problem lies in a key variable, an assumption built into the algorithm, or the project’s general approach—the pilot just does not perform as expected.

Piloting an analytics project, like any effort to innovate in the public sector, is somewhat at odds with the bureaucratic preference for consistency and risk avoidance, but it is a critical phase that can yield important insights for improving performance when it is time for implementation on a larger scale. Moreover, starting with small-scale pilots can help limit risk and demonstrate clear results.

Piloting also allows for much needed course corrections to help better transition efforts in project-scoping to implementation; adjusting project parameters during the pilot phase can increase the likelihood of success at implementation and beyond.

5. Implement and Scale the Model

Research and literature on implementing and scaling analytics projects remain limited, and given the variability of structures, budgets, and objectives for analytics projects, identifying generalizable practices for scaling these projects is challenging. There is no cookie-cutter solution for scaling up, and the key is to iterate, iterate, iterate.

One of the biggest mistakes that people make with regard to data analytics is thinking that once a model is built and goes live, it will continue working as normal indefinitely. On the contrary, models will actually degrade in quality over time if they’re not continuously improved and fed new data. In order for it to remain useful and accurate, you need to constantly re-evaluate, retrain it, and develop new features. If there’s anything you take away from these fundamental steps in analytics and data science, it is that a data scientist’s job is never really done, but that’s what makes working with data all the more fascinating!

DATA
10001
10100
10001
101011
10010
010111
65.7843
22.106
R - 7.4
P - 8931
VE - 0084
Q / U - 86
J _ FR - 7854 - 8923
24.063
F - 2.4

Chapter Five

05

How to Present Data Ethically

Dos and don'ts for presenting data ethically

We know by now that data presentations can enhance people's understanding of complex information, but it's important to acknowledge the ways that they can mislead audiences, too. Hiding relevant numbers, presenting too much information, or choosing the wrong kind of graph risks unintentionally distorting the message in your data.

While making mistakes is an inevitable part of learning how to visualise data, there are some ground rules you can follow to help make your presentation as accurate and honest as possible.

We've listed four rules below:

Don't let confirmation bias dictate your choices

Humans are confirmation machines. Without knowing it, we can easily become laser focused on finding only the numbers that support what we believe. When we want an idea to be true, we tend to seek evidence that supports our beliefs and interpret it in a way that confirms these beliefs. Being aware of this fact is a good place to start when you're thinking about conveying data to an audience.

Creating a story or a visualisation out of data may devolve into a falsification process sometimes if you only seek the numbers that support your argument and nothing beyond that. So how can we counteract this? You must train your brain to do the opposite: to expand your scope and seek out information that complicates your ideas. Numeracy is like riding a bike: practice paying more attention and it'll become second nature eventually.

Don't conceal uncertainty

Uncertainty is an inherent fixture of life. Don't deprive your audience of nuance, and be sure to include data points that contradict your argument. Uncertainty is essential, and you cannot hide it. If you withhold information that casts uncertainty on your argument, you are manipulating the existing data to create a narrow picture of what's happening.

This is also about creating a space for those stimulating conversations that may arise from uncertainty. A good chart may help you to answer a question ... but they can also be great for piquing our curiosity and prompting us to ask more questions.

Revealing contradictory or confusing data points also shows integrity and can increase your credibility in the eyes of your audience. So, don't be discouraged if your message is complicated by ambiguous data – embrace it as a way both to spark conversation and establish trust with your audience.

However, it's still important not to overload the audience with extraneous information. Use your best judgment and ask yourself: "if I was listening to this presentation, would excluding this information help or hinder my understanding of the core issue?"

Do contextualise your visualization

To mitigate potential misunderstandings, contextualise the data that you're presenting thoroughly. Not only will this minimise the chance of your audience misinterpreting your data, but it will also help you to keep your own confirmation bias in check. Lay out your case step by step and attach each link of your reasoning chain to preceding and subsequent ones. Do this and it'll be much harder to make unfounded and deceptive claims.

Clarify the origins of the data and offer possible alternative interpretations when you can. This is an essential practice when you're presenting data – we want information to be presented simply, but never at the expense of understanding.

Do use reliable data

It might seem obvious, but it always bears repeating: use data from trusted and credible sources.

There are a few ways to ensure that your data is accurate. One of the easiest is to make sure that your data is coming from a primary source. If you find data you'd like to use, track down the original source and verify that it is legitimate. Ask under what conditions the data you're using was generated. Was it published by an organisation with a political agenda? Is it biased? Use discretion if the answer is yes.

Another way to increase data reliability is to use the most recently published data available: if it's older than a year or two, make this clear in your presentation.

And finally, if you are unsure about your process, collaborate with experts! This is especially true if you are making your data-driven argument public. Never, ever publish something that hasn't been looked at by an expert.

Chapter Six

06

Key Challenges of Data Analytics in the Public Sector

Resource > Your Chart
Business Chart - Visual
Business Chart
5M$
4M$
3M$
2M$
1M$
0M$
2
3
4
5
6
7
8
Who is your audience and what are their needs? This can help you better articulate the benefits of doing business with you and deliver a smarter product or service.
Interactive User
1,505
New Users Registration
18,321
Registered Users
Realtime Dashboard
37.91%
31.86%
30.23%
Focus on Your Finances
Whether you're a full time, part time or freelance photographer - or even if you make a few bucks on the side from your photography - you are your own business.
Target
Space Usage (750 Mb)
375 Mb
used
250 Mb
available

Despite the importance of data analytics, the public sector in many countries is not yet ready to mine value from public data. There are several barriers that prevent the sector from realizing the potential gains of data analytics. Below are some foundations of the effective use of data analytics which are sometimes neglected by the public sector.

As data work increasingly fans out across agencies at all levels of government, public servants who do the actual work are starting to identify a set of shared challenges.

IT and innovation department staffers, as well as technologists who work within agencies like health or public advocacy, are using datasets, analytics and evidence-driven governance techniques at an increasing rate in government. This isn't a new revelation. What is, perhaps, less commonly acknowledged is that staffers doing this data work are facing a set of challenges that are the same or similar, whether they work in city hall, or the national government.

1. The cost of data visualization products is often a difficult barrier to overcome. Data analytics would be easier if organizations could afford high-end data visualization software, but the barrier of cost to invest in that is so high, that organizations instead mostly find workarounds that are often cheaper in the short term but ultimately more costly in the long run.

2. The inherent nature of budgeting creates a unique financial challenge for government technologists at all levels as well, at least relative to their colleagues in the private sector. Government budgeting is done considerably further in advance, which means that technologists who work with data-driven pilot projects must commit in advance to requests for funding.

3. Another challenge is recruiting qualified talent and finding existing talent already working within government that possesses data skills. There are many new tools available allowing the public sector to analyze data. For example, SAS and R programming are common tools in statistical analysis and data modeling, while Tableau public and Python are widely utilized for data visualization. Although data analytics can be carried out by in-house data scientists and external experts, public-sector employees must possess basic

22
21
20
19
18
17
A
B
C
plan
Relative activity
product A
product B
product C
Relative activity of subdivitions of main

knowledge of the process, in which statistics is a core discipline. Without knowing how data are collected, analyzed, interpreted, and visualized, their capacity to extract useful insights for better decision making cannot be ensured by merely observing the results from data analytics.

4. Data analytics start with data collection. As the public sector carries out various functions, it has ample access to diverse sources of data. Despite the potential for data collection, the public sector at times does not pay attention to the standardization of data, for which a common format and structure are necessary. If there is no data standardization, it is difficult to integrate data and extract valuable information from it.

In addition, collected data should be digitized and stored at an organizational as well as national level. To do this, as many private organizations do, the public sector needs to appoint a chief data officer (CDO) and establish a team dedicated to managing data quality. Collected data should be open to public access. Data release not only increases the transparency of the operations of the public sector, but also creates new opportunities for improvement.

5. The public sector must realize that data are an important asset to design better policies and implement them more effectively. To take advantage of this, data collected by multiple units and organizations should be unified through cleansing, mapping, and transformation. However, data in many cases are fragmented and spread out within an organization and across organizations, which hinders effective use. Among the reasons for data fragmentation lies a perception that data and information are in the power of data holders and seen as proprietary, not to be shared. Both organizational leadership and a CDO should make efforts to eliminate this antiquated perception and to create a new culture of data sharing.

6. New environments, led by big data, are revolutionizing the volume, variety, and velocity (aka the 3Vs) of data. Accordingly, new tools have been developed to assist organizations to conduct data analytics easier and faster. One of the prerequisites that enable the public sector to incorporate these new analytical techniques successfully into its organizations is continuous investment in the capacity for data analytics among employees. Building data analytic capacity is not restricted only to the learning of fast-changing analytical tools. It includes a clear vision that aligns data analytics with organizational mission, basic knowledge of statistics, and data ethics.