DATA ANALYTICS UNIT (DAU)

DAU PROCESS DOCUMENT

DAU, APCFSS, Finance Department, Government of Andhra Pradesh

Table of Contents

The Government of Andhra Pradesh has established a Data Analytics Unit(DAU) as part of its commitment to encourage the use of evidence to inform policy formulation and program implementation.

The DAU would undertake analysis of administrative and other data to provide actionable and relevant decision-support information for senior officials and field functionaries. It shall function as a shared service available to state government departments on demand for building capabilities within Departments on data analysis and evidence-based decision-making.

This document lays down the objectives, processes, and personnel of DAU and the way in which it will work with Government Departments.

WHAT IS DAU?

The DAU is currently a dedicated and small centralized unit of data scientists within the Finance Department that are working with other government departments to co-develop data analysis requirements, develop frameworks and methods to analyse data to assist in decision making, and eventually train and capacitate in-house departmental teams for data analysis and support the department through the process.

The team is composed of professionals employed by the Government of Andhra Pradesh with ability to handle large data sets, code in different languages, and leadership that has experience of working with government departments to understand and fulfil their needs. For specific needs and for their own capacity building, they partner with other organizations, who have the requisite experience and expertise.

WHY IS IT NEEDED?

In the past decades, there have been two trends. First, linked, in part to Aadhaar, there has been significant digitization of government processes and databases, leading to a lot of administratively generated data, but so far, limited analysis. Second, computing has become inexpensive, due to fall in equipment prices and technologies like cloud computing. This advancement has made it possible to tap the profound potential of very large datasets inexpensively to discover deeper insights.

Data is now available at very granular scale, for both businesses and individuals. There is data on tax payments at the level of invoices, data on links between businesses, data on school achievement of students, on savings behaviour of SHG members, on cash transfers to beneficiaries, a growing method of delivering social support. In Andhra Pradesh, the Gram Sachivalayam and Ward Secretariat (GSWS) database used for the

six-step verification for government schemes covers almost the entire population of the state and can be used as a common benchmark for various other data sources. Organizations like APSSAAT are building the capacity to quickly implement large state-wide surveys to supplement administrative data.

Current products like dashboards merely describe data and are limited in their use of data and effectiveness as decision support tools. The question is whether more effective use can be made of this mass of data to support departmental decision-making, by using more sophisticated statistical models and developing a more structured process driven approach to collecting, organizing, analyzing and using data. This document outlines such an approach.

Privacy and Data: As large data is harnessed in a distributed fashion to ensure maximum effectiveness, the risks of breach of privacy for both individual and firms grow. A culture of privacy and non-disclosure of identifiable information must be inculcated in all those who work extensively with data. DAU is committed to developing this culture.

National Impetus

At the national level too, there is a strong impetus towards data driven governance, with the NIC (National Informatics Centre) playing a key role (see Box 1). However, the maximum benefit from such initiatives can only be derived when there is sufficient local capacity to absorb and use these tools and a willingness to modify decision processes and data flow to optimize the benefits of data-supported decision making. It is only then that a state can begin to realize the "huge untapped potential" of data.

Box 1: Data Driven Government according to NIC

In Government, the most traditional use of data analysis has been the statistical analysis of data collected through various surveys, census, indices, etc.… With the launch of the Digital India Program,…a wide range of government initiatives and schemes (Central as well as State) are today making extensive use of data and … [r]ight from the concept to formulation, implementation to the monitoring of a scheme, data is now being extensively used in almost every aspect of a project or initiative. For instance, data is at the core of many flagship programmes such as Swachh Bharat Mission, Housing for all, One Nation One Ration Card, Pradhan Mantri Ujjwala Yojana, Fertilizers Distribution, to name a few. NIC's in-house developed tool called Darpan has helped many ministries and

organizations by extracting data from various IT systems and create dashboards and insights from this data.

[D]ata can prove to be highly useful for the formulation of poverty alleviation schemes as well as subsidy distribution schemes. Various schemes of the government such as the MGNREGA, Pensions Scheme, Farmers Subsidy, Benefits for unorganized labor, Scholarships, etc. can make use of data analytics to identify the right beneficiary, understand their socio-economic status, and use technology solutions for timely dissemination of benefits, etc. These programmes are touching the lives of millions of citizens in India and thus ensuring equitable and rightful distribution of benefits.

Similarly, fields like Criminal Justice and Judiciary can consume data to analyze crime patterns, locate the criminal networks and hotspots of potential crimes, etc. This would help the authorities take corrective measures and prevent any such incidents from happening. Data is extremely valuable in fraud prevention also. Many financial systems today are employing data to detect fraudulent activities and it is now suggested to integrate a fraud detection module while setting up any financial system.

During the COVID-19 pandemic, data has been extensively used for contact tracing, prediction of hotspots, trends analysis, and take appropriate measures to curb the spread of the virus. Data was also used for the management of hospitals and the supply of essential medicines and essential goods to citizens at large.

However, various researches also suggest that ***much of the data is still not analyzed and has huge untapped potential.*** *One of the major challenges right now is the fact that data is currently residing in silos and thus to unleash the true potential of this data, various IT systems must collaborate and operate in a symbiotic fashion. The National API Exchange Platform set up by NIC is supporting a safe and secure flow of data between different systems, basis the mutual understanding of the participating entities. The need of the hour is to understand the significance of good quality data and employ it for the betterment of society.*

Source: Neeta Verma, Director General, National Informatics Centre
https://www.nic.in/blogs/data-driven-government/

The value proposition of data analysis depends on the actionability of its outputs. Some outputs of data analysis can directly be the basis for administrative actions, whereas some others would need to undergo further analysis before they can be acted upon. The degree of actionability also varies depending on the administrative leve

The DAU will follow the following framework for actionability:

Primary Insights: Directly Actionable Insights:

These insights are immediately actionable and can be used by officials for administrative decisions without further analysis. They provide granular signals or anomalies. For example, in Commercial Taxes outliers among taxpayers with significant deviation on a parameter could immediately trigger some enforcement action (like audits or inspections) on those individuals. The data analysis output offers actionable information that can directly inform decision-making and enforcement actions. Such information is most useful for frontline officials.

Secondary Insights: Requiring Further Scrutiny and Examination

These insights necessitate additional examination or investigation by departmental officials for them to become actionable. The insight raises concerns or potential issues, but more in-depth analysis is required for administrative actions. For example, in commercial tax we have sectors or regions with suspiciously large Input Tax Credit (ITC) claims or excessive Electronic Way Bills (EWBs) generation that needs closer scrutiny. The data analysis output should ideally offer specific leads or directions for inquiry, aiding officials in conducting more thorough investigations.

Tertiary Insights: Demanding Deeper Analysis

These insights indicate important or troubling patterns but demand additional analysis by the DAU before becoming actionable. For instance, trends in Integrated Goods and Services Tax (IGST) inflows or ITC claims might suggest evasion, yet they require deeper analysis to uncover sectors, geographies, and taxpayers who can be targeted with enforcement actions. Sometimes the trend might be a reflection of underlying economic factors and therefore may not demand action by the Departmental officials. The data analysis output would highlight broad signals, and push DART into doing comprehensive analysis to reach a stage where the insights can be directly acted upon.

By using this framework, the DAU can present its findings in a structured manner, allowing officials to quickly understand the nature of the insights and the appropriate level of action required. This promotes a focused approach to decision-making and

enforcement actions, ensuring that the department's resources are utilized efficiently and effectively based on the actionable nature of the data analysis outputs.

WHAT WE DO?

DAU will do the following:

(a) Requirement Identification :

Objective:
Work with departments and units within the Government of Andhra Pradesh (GoAP) to identify areas where data can facilitate better decision-making.

Activities:
Collaborate with various departments and units to pinpoint their specific needs.

Analyze what kind of data can satisfy those needs and aid in decision-making.

(b)Data Integration:

Objective:
Layer internal and external data sources to create a comprehensive dataset that can be a reliable foundation for analysis.

Activities:
Assist departments in cleaning and systematizing their data.

Identify and integrate additional public data or data from other departments within GoAP to enrich the dataset.

Develop frameworks to analyze the integrated data effectively.

(c) Real Data Analytics:

Objective:
Delve deep into the data, going beyond mere dashboarding to extract valuable information, trends, and correlations that can guide decisions.

Activities:
Analyze data to identify useful patterns and trends.

Extract valuable insights from the data to aid in decision-making.

(d)Analytical Instruments:

Objective:
Develop advanced data-centric models ranging from foundational tools to AI-driven solutions to support decision-making.

Activities:
Create frameworks that can identify useful patterns in the data.

Develop sophisticated statistical models and spatial representations.

Utilize machine learning techniques to analyze very large datasets.

(e) Decision Support:

Objective:

Offer insights derived from data analysis to help make informed choices.

Activities:
Generate and communicate support materials that are easy to understand to aid in decision-making.

Work closely with departments to develop their capacity to conduct and interpret data analysis.

(f) Data Visualization:

Objective:
Enhance decision support by presenting data in intuitive visual formats for a clearer understanding and to facilitate actionable insights.

Activities:
Develop tools that present data in visually intuitive formats.

Assist in transforming complex data into actionable insights through effective visualization techniques.

End-to-End Data Analytics and Decision-Making Framework

The various stages of data analytics will be as follows:

1. **Requirement Identification and Analysis:** Partner with GoAP departments to co-create datacentric approaches for analytics driven decision making.
2. **Data Acquisition:** This involves gathering relevant data from both primary and secondary sources both within a single department or across departments. The

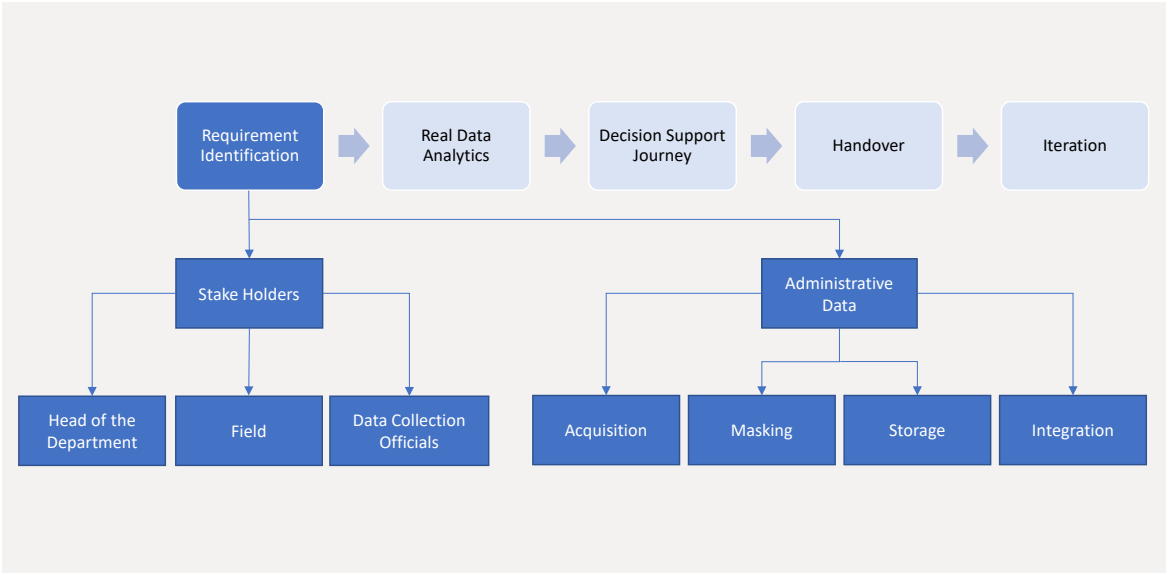
various sources may include databases, APIs, websites, censuses, or surveys.

3. **Data Storage:** The data across departments may be available in multiple different structures and formats. We would be storing them in a structured and accessible format, often in databases, data lakes, or data warehouses to ensure uniformity and one source for analysis.
4. **Data Cleaning and Integration:** Raw data is cleaned, validated, and transformed to ensure consistency and accuracy. Different data sources are integrated to create a unified dataset.
5. **Data Analysis:** This stage involves applying statistical techniques, mathematical models, algorithms, and visualization tools to extract insights and identify patterns, trends, or relationships within the data.
6. **Data Modeling:** Creating mathematical or statistical models to represent and understand complex systems or phenomena based on the analyzed data. This may involve machine learning, predictive modeling, or optimization techniques.
7. **Data Visualization:** Presenting the analyzed data in a visual format, such as GIS-based maps, charts, graphs, or dashboards, to facilitate understanding and decision-making.
8. **Reporting and Presentation:** Communicating the findings and insights derived from the analysis to stakeholders or decision-makers through reports, presentations, or interactive tools.
9. **Decision Making:** Using analytics insights to make informed decisions, optimize processes, allocate resources, or solve problems within an organization or domain.
10. **Implementation and Execution:** Putting the decisions into action, implementing recommended strategies, and monitoring their impact on the desired outcomes.
11. **Performance Measurement:** Tracking and evaluating the outcomes and performance of the implemented strategies or actions to assess their effectiveness and identify areas for improvement.
12. **Feedback and Iteration:** Incorporating feedback from the performance measurement stage to refine and improve the analytics process, models, or strategies for future iterations.
13. **Continuous Improvement:** Iteratively refining and enhancing the analytics capabilities, infrastructure, and methodologies to stay up to date with evolving technologies, data sources, and business requirements.

HOW DO WE OPERATE?

In this section, we go through each of the aforesaid stages in some detail, explaining their respective processes.

Figure 1: Process Map: Requirement Identification



Design Workshops for Need Identification

There are three broad steps in developing a framework to ask purposefull questions

a. Decision support needs
b. Data needs.
c. Stakeholder engagement and conceptual approval for Data-Driven Decision Support

Decision Support Need

The first step in using data to assist with decision-making is to understand the key decision-makers and the nature of decisions they need to take. Some of these can be monito ring, others can be action related, depending on the level of the decision-maker. For example, the head of department or the secretary level officer, in a tax function may be in interested in monitoring the overall system along key metrics (e.g., revenue growth) and summary measures of efficiency (e.g., tax buoyancy). At the same time, the officers supervising the field may be interested in process compliance and identification of leakages, e.g., using data to identifying specific retail outlets or dealers that require field-level investigation because they are exhibiting anomalous behaviour.

Data Need

The second step is to understand what kind of data is being collected and available to support decision-making. The persons in charge of collecting and organizing the data are key stakeholders in this process.

Stakeholder Engagement and Conceptual approval for Data-Driven Decision Support

The third step is to determine how the data can be analysed to support decisions that have been identified in the first step and get the conceptual buy-in of the stakeholders, both in the generation and availability of the data and in the usage of the results from the analysis as a decision support tool.

Administrative data is typically not generated to test a hypothesis or answer a question. It is typically generated as a form of "digital bookkeeping", as a digital summary or footprint of existing government processes. One must work in reverse to understand what kinds of questions can be meaningfully answered given data availability and quality.

In order to understand the nature of data that is available, it is necessary to organize the following activities, viz.:

1. **Design Workshop:** DAU will organise a workshop to delve deep into the critical aspects of decision support and data needs to enhance decision-making processes at various organizational levels. Initially, we will identify the primary decision-makers and understand the nature of the decisions they are required to make, which can range from monitoring tasks to action-oriented decisions. We will draw examples from different hierarchical levels, illustrating how data can aid in monitoring key metrics such as revenue growth and identifying operational inefficiencies. Following this, we will shift our focus to understanding the current data collection processes and the types of data available to support informed decision-making. Participants

of involving key stakeholders responsible for data collection and organization. The workshop will foster discussions on how to analyze the available data effectively to support the identified decisions and secure stakeholder buy-in conceptually.

2. **Data Integration:** The journey of building a comprehensive data set is as follows
    1. **Data Acquisition** This would need mapping of available data sources, viz.:
        a. Administrative data being collected within the department, e.g., data from tax forms filed by taxpayers or sales of GROs (retail liquor shops), etc.
        b. Administrative data being collected in other departments that may be relevant for the department seeking to use data for decision support, e.g. the number of metered households collected by the electricity department may be useful for the property tax collection in Municipal Administration department
        c. Public data, e.g. census data, large survey data, etc.
        d. The various sources may include databases, APIs, websites, censuses, or surveys.
    2. **Data Storage:** The data from the various sources may be available in multiple different structures and formats. The DAU will need to resolve inconsistencies between different sources, and store the rationalised data in a structured and accessible format to ensure uniformity and availability for analysis, with proper security and provenance (i.e., it should be possible to identify the source for all the data in the database).
    3. **Data Cleaning and Integration:** The data from the various sources are to be cleaned, e.g., some information may be in incorrect columns, such as the state name may show up in the district column, validated, e.g., items which should match across databases should actually match, and then transformed to ensure consistency and accuracy, so that they can be read by computer programs analysing the data. This pre-processing of all data sources ensures uniformity and proper linkages across data sources.
    4. **Data Masking:** Finally, if the data is to be shared, e.g., with third party experts for analysis, it would need to be masked, i.e., the identifiers that can identify an individual or firm have to be replaced with unique (in order to preserve the ability to do analysis) but non-identifiable alphanumeric strings.

After these steps, the data would be in a state to be meaningfully analyzed, and as per the third step, various options for using the data to support the decisions as identified by the department can be worked out in a consultative manner. The key outputs would be:

1) Requirement Analysis and Deliverables Document
2) Processed Data in Usable Form
3) Project Plan with Work Items and Task Plan

Box 2: Example of Design Workshops for Need Identification

The Excise Department's need was to establish a reasonable basis for fixing a target, i.e., expected revenue that should should have accrue at the level of the mandal and GRO. The target could have been either the number of bottles sold or total sale value. However, the number of bottles was an inconsistent measure because price and supply of various varieties were being constantly updated. So, a detailed workshop was held by bringing in all the stakehoders to conclude that we should be targeting the daily sales value, aggregated across all varieties, at the level of the GRO.

Real Data Analytics

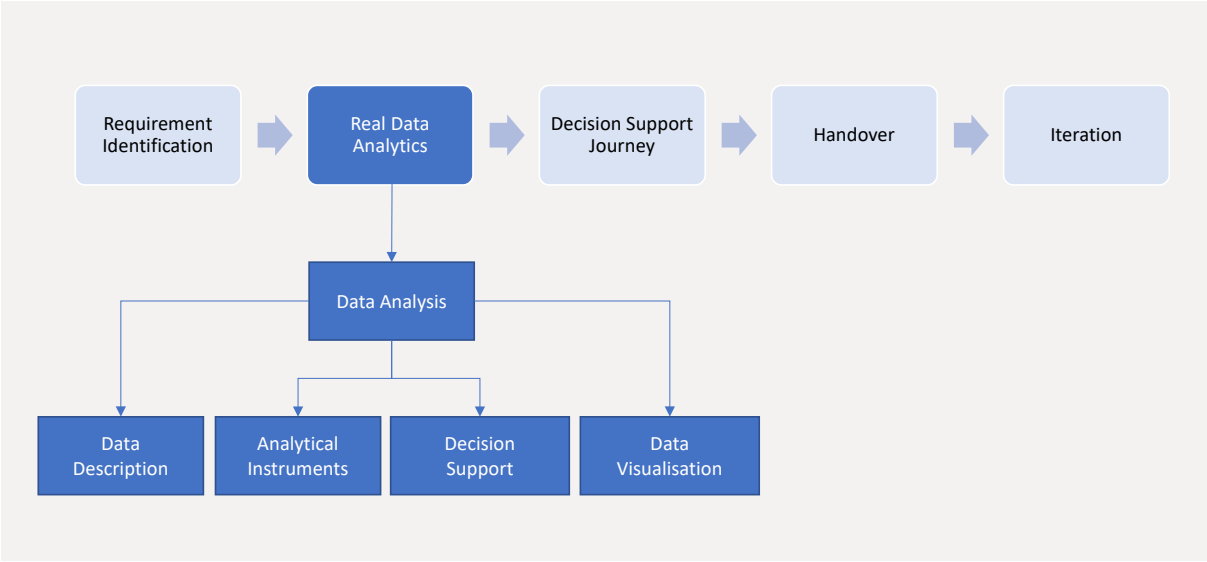
Once there is clarity about the need and the data that would be used, the next task is to decide about the kind of analysis that could be applied to the data.

A simple-minded example of such analysis from Box 2 above would be as follows: *Calculate the average sales value of all GROs in the state and identify all GROs with sales below 80% of the state average as possibly under reporting sales*. This is obviously inappropriate since the GROs in less populated or lower income areas would have lower sales and vice versa. An improvement on this method would be to have the same method, but use the district average instead of the state average, but that would still leave much to be desired.

Thus, of the key goals of a DAU is to build core competency in itself and in the departments about the appropriate method for analyzing data, often using advanced analytics, through use-cases, specialized skills, and inclusion of data that are available across departments.

This would require an understanding of the data and familiarity with statistical tools to analyse these datasets. As the DAU conducts more analyses, it will generate, in collaboration with different departments, a corpus of use-cases, with different methods of analysis and a complementary set of analytical skills along with properly catalogued data. Along with access to the data, the capacity to use these methods and the associated skills will be built by DAU in the user departments. The methods of analysis can be broadly grouped into three types:

Figure 2: Process Map: Real Data Analytics



1. **Data Description:** Data description is an essential step in data analysis as it helps to understand the data. It refers to the process of summarizing and presenting the key characteristics and properties of a dataset, i.e., an overview of the data, including its structure, content, and relevant statistics. This enables a first-cut identification of any patterns or anomalies, and helps to make informed decisions. One key method is to filter the data, i.e., compare summaries of key pre-identified sub-populations, e.g., revenue performance of different mandals in the state.

2. **Analytical Instruments:** However, in many cases, this kind of descriptive analysis is insufficient for the purpose and the analytical need requires the use of more sophisticated statistical models and computational algorithms.

3. **Decision Support:** Typically, the aim is to try to understand data patterns of a key metric, such as sales of a GRO or share of GST paid in cash. At this point, two kinds of analysis can be carried out. The first is to characterize sub-populations have systematically high/low values on the metric. For example, we may want to understand which industries pay more of their commercial tax in cash vis-à-vis input credits, because, for instance, they have a higher share of value added. The second is to characterize individual "outliers". In the above example, we may want to characterize taxpayers that are paying an unusually low proportion of their taxes in cash as the second type of analysis.

   Statistical models can combine such different kinds of analysis, along with accounting for multiple factors that could affect the value of the metric. In this example, apart from the industry, it could include, inter alia, the size, age and location of the taxpayer.

4. **Data Visualisation:** Since many of the administrative structures in India are spatial, e.g., districts, mandals, circles, urban local bodies, wards, etc., it is useful to organize data in a spatial manner for purposes of monitoring and oversight.

   GIS analysis is designed to visualize and analyze geospatial data in a user-friendly and interactive manner. It provides a comprehensive view of spatial information and allow users to explore, analyze, and intuitively compare similarly placed regions through a common set of key performance indicators. One can then track regional performance and identify outliers, as above, by spatial/regional factors.

   In addition to such depictions, certain types of statistical models use the spatial information in the data as an explanatory variable, such as proximity of firms to each other, location in the same industrial estate, etc.

In this phase, the key outputs, thus are:

1) Analysis according to the Requirement Analysis and Deliverables Document

2) Key Insights that must be communicated to stakeholders

Box 3: Example of Real Data Analytics

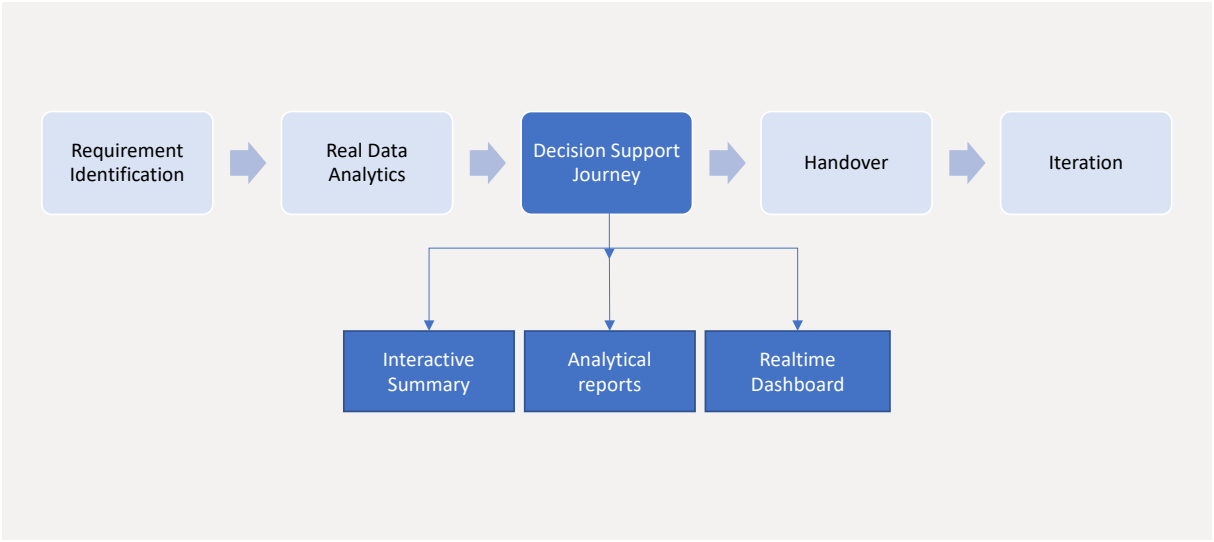
The data on daily sales value of GROs provided by the Excise Department had few predictors that could be used. A team from a third party expert organization (Centre for Policy Research) were asked to develop the core statistical model. Together with them, DAU decided to use a "fixed effects" approach that controlled for factors at the level of the GRO (by incorporating the average of the previous year's sales by the GRO as an explanatory variable) and effects related to specific days, e.g., weekends, festival days, etc. by using day-specific dummies. Since this approach accounted for factors specific to individual GROs by design, idiosyncratic explanations provided by field level officials to explain differences in performance across GROs were already incorporated. This addressed a key concern of an important stakeholder. The model was coded in R by CPR and produced 'Expected Sales per day' as a metric to determine performance of GROs and spatial units like mandals. This model was then replicated in Python by the DAU. DAU than used a comparison metrics of 'Expected Sales per day' with 'Actual Sales per day' to define a benchmark of performance for each GRO and Mandal. Thereby identifying the exceeding and lagging GROs, Mandals and Districts within AP. DAU then prepared historical and live performance dashboards suited to help the department make data driven decisions.

Decision Support:

The Journey from Analytical Instruments to Intuitive Visual Formats

This is the critical aspect of the process. In order for the execution to improve decision making, the results must be clearly communicated to the decision maker in a manner that can support her or his decisions. This can broadly comprise three types of outputs, viz. analytical reports, interactive summaries, and real-time dashboards. Before describing them, it is useful to review the characteristics of data and how the nature of data is related to the type of output.

Figure 3: Process Map: Decision Support Journey



First, it is necessary to determine whether the dataset is static or dynamic (this is not to be confused with common academic terminology which uses "dynamic" to refer to any data that includes a time component). A static dataset is one in where the data that is to be analysed will not be change or be updated during the analysis, e.g., data on tax revenue for the previous fiscal year. Conversely, a dynamic dataset is one in where the data therein is being frequently updated, e.g., real-time tax receipts.

Table 1: Nature of Data and Analysis

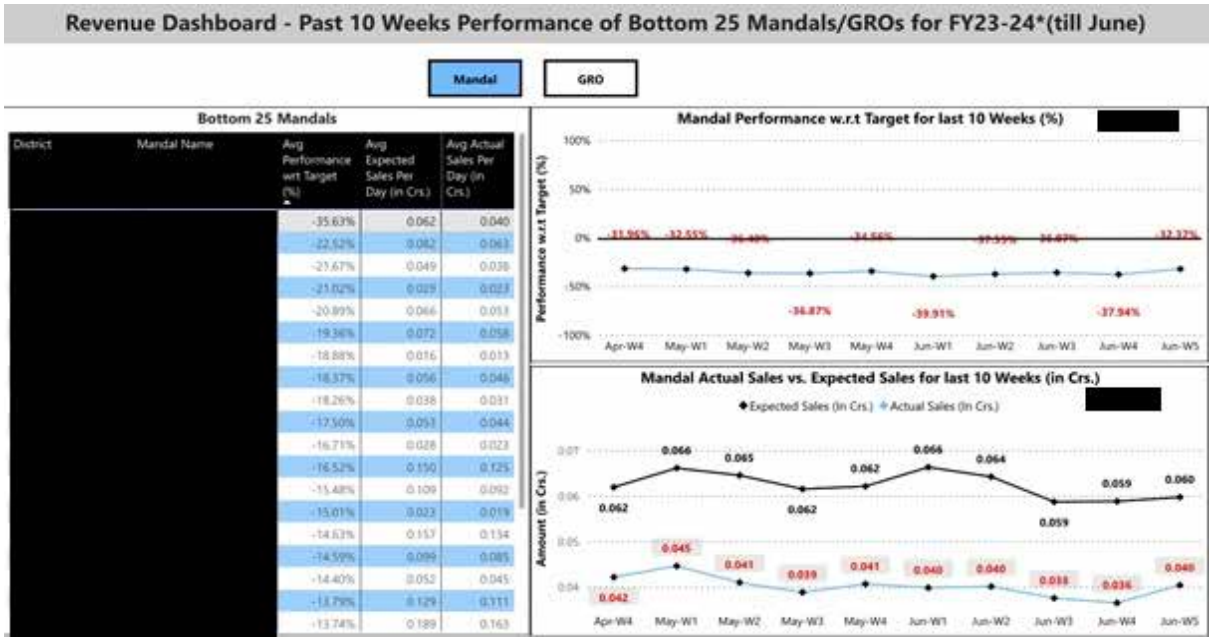
	Low Complexity	High Complexity
Static Data	Interactive Summary	Analytical report
Dynamic Data	Real Time Dashboard	Full time monitoring

Second, it is necessary to understand the how the data are to be used. Often, the best decision support allows the decision-maker to visualize the details of the data in a comprehensible and interactive manner. Here, there is low complexity in the analysis, as this is fundamentally about parsing a small number of internally consistent data variables. At times, we are interested in testing more difficult hypotheses that require using data across multiple sources, and which must be subjected to rigorous statistical tests. Here there is high complexity in the analysis. Based on these considerations and the type of dataset and the complexity in the analysis (see Table 1), outputs can be classified into three types:

1. **Interactive Summary:** While the data are static this enables the decision-maker to "dig deeply" into the data in an interactive manner. Rather than trying to understand the relative predictive power of various factors, the key goal of such an output is to allow the decision-maker to understand the full data to identify areas for administrative action and combine his or her domain knowledge with data patterns. Often, this kind of interaction can help to identify predictive explanatory variables.
2. **Analytical Report:** This is based on more sophisticated statistical models that seeks to identify the relative importance of different variables in explaining various outcomes of interest. This requires cross-referencing across multiple datasets, significant data processing, and rigorous statistical tests. The goal of this output is to communicate the statistical findings to the decision-maker in a manner that will help him or her use it as a decision support tool.
3. **Real-Time Dashboard.** This kind of output is designed to support the decision-maker to make real-time decisions more effectively. As the data are dynamic, the dashboard constantly recalculates and visualises previously agreed upon metrics, thus supporting real-time monitoring.

In Figure 4a, the revenue raised from the 25 lowest ranked mandals (in terms of revenue) is presented. In one segment of the screen the names of the mandals (masked in this graphic), their performance (as % of the target), the target revenue per day and actual revenue per day is presented. On other segments of the screen, the same three metrics are presented for a mandal selected by the decision maker (masked in this graphic), summarized by week, for the previous ten weeks. This permits the decision maker to monitor the performance of the mandal over the recent past, to see if it is improving or deteriorating. In this instant case, it appears as if the performance in the past ten weeks has been stable, but at a level much below the target.

Figure 4a: Example of a mandal level dashboard



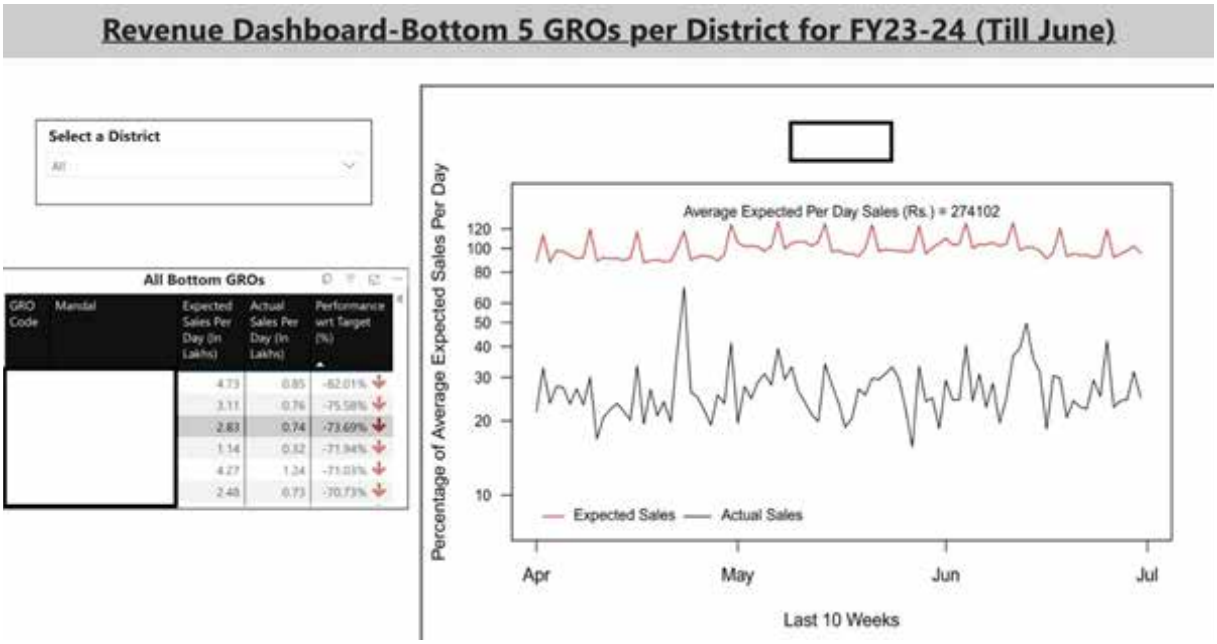
In Figure 4b, one can see a GRO level dashboard. In one segment of the screen, the bottom GROs in a given district are presented, with one GRO highlighted. The right hand segment shows the daily performance over the past ten weeks for the highlighted GRO (number masked in this case) along with the *modelling exercise, which controls for various factors to predict the daily sale of alcohol*. This generates the red line i.e., target sales (red line). It shows that the performance of this particular GRO had been steady but much below target over April to June, indicating that the issue is not a one-off decline, but a record of chronic underperformance.

In both Figure 4a and 4b, as new data arrives the dashboard is updated automatically – the earliest week/day is dropped and the latest week/day is added.

Where is the use of statistical modelling in this exercise? It is not immediately obvious but the target daily sales by the GRO is the outcome of a statistical modelling exercise, which controls for various factors to predict the daily sale of alcohol. This generates the red line in Figure 4b and also the deviations from the target in Figure 4a above.

In the decision support phase, the key output (which may have many components), is thus *Deliverables to stakeholder, consistent with agreed solution deliverables.*

Figure 4b: Example of a dashboard-GRO level



Box 4: Example of Decision Support Journey

For the Excise Department, DAU has created a series of dashboards and interactive summaries to communicate the results of the analysis in a manner that can support decision making, as shown in Figures 4a and 4b. In particular, from the dashboards and summaries, the decision maker can immediately select and display the "outliers' which show particularly poor revenue performance from the sale of liquor and take appropriate action. The key here is the benchmark vis-à-vis which the outlier is identified. At the field level, the tool identifies the GROs that need to be investigated and asks the officer to determine reasons in writing for poor performance of the GRO. This feedback from the field will help to refine the metrics to select outliers in the next iteration of the model.

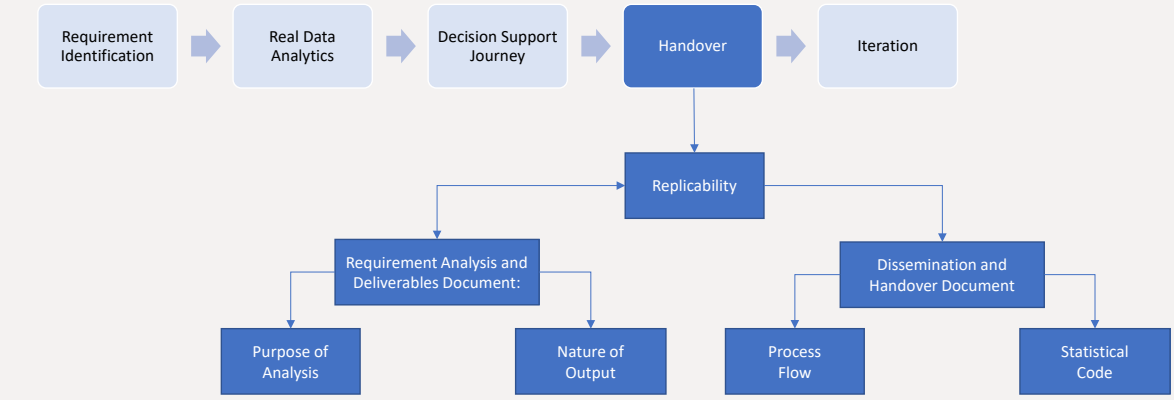
Handover

During the prior phases, the Data Analytics Unit (DAU) will work hand-in-hand with various departments to develop a rich portfolio of use-cases, each showcasing different analytical methodologies and well-organized data.

The primary goal of Handover phase is to foster skill development and capacity building within specialized units of the user departments. To ensure the success of this initiative, it is imperative to meticulously document each step involved in determining the specific statistical analysis approach and the subsequent calculations performed. This approach guarantees that the analysis can be replicated by analysts who may not have prior knowledge of the datasets, calculations, and algorithms utilized initially.

By adhering to this stringent documentation process, we aim to create a decision support tool that stands up to scrutiny and can be reliably used to facilitate informed decision-making. This handover process is designed to be a cornerstone in establishing a credible and replicable analytical framework that can serve the various departments efficiently and effectively.

Figure 5: Process Map: Handover



To generate replicable protocols, we require three forms of documentation, viz. the core guiding document, the associated process flow and the accompanying statistical code.

1. **Requirement Analysis and Deliverables Document:** The guiding document states the purpose of the analysis and describes the nature of outputs, e.g., dashboards, model metrics, etc. It also describes the data and the algorithms used in generation of these outputs. In particular, the data section includes the provenance of various data-sets, the variables included in each dataset (including any transformed variables, e.g., interpolated values or aggregated values, and the reasons for doing so) and a detailed

codebook for the data. The algorithm section includes an exposition of all algorithms used for the dashboard and/or analysis, including the associated statistical and mathematical properties.

2. **Dissemination and Handover Document:** The Dissemination and Handover Document serves as a comprehensive guide detailing each step involved in the analytical process, including the specific data and statistical code utilized at each juncture. It outlines the data storage solutions adhering to the highest standards of data privacy and delineates the pathway to access the stored data. This document is instrumental in facilitating the creation of final dashboards and model metrics, clearly describing any intermediate outputs required in the process. Moreover, it houses the statistical code responsible for generating the core metrics, algorithms, and analyses. This code is thoroughly commented to elucidate the intermediate steps and justify each coding decision, ensuring a smooth handover to analysts unfamiliar with the code. It also specifies the storage locations of the data files used, requiring timely updates if there are any changes to the file locations to maintain the integrity and functionality of the analytical process. This meticulous documentation aims to foster a transparent and replicable analytical environment, encouraging informed decision-making through a deeper understanding of the underlying processes.

These two key pieces of documentation allow outputs to be replicable. Replicability is important, as it allows for the rapid correction of errors. It also maintains consistency of outputs even if there is turnover of personnel and also provides an efficient way to quickly adapt and extend existing analysis.

Box 5: Example of Handover

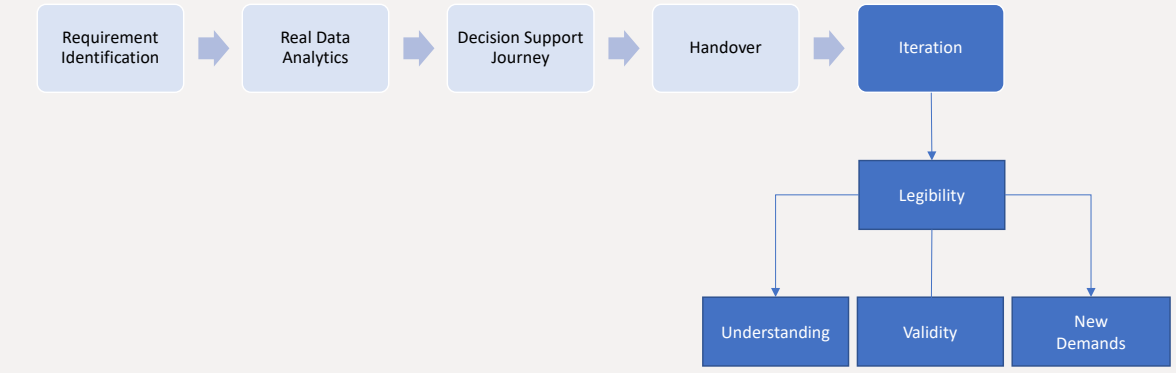
For the Excise Department, together with CPR, the DAU has already developed Requirement Analysis and Deliverables Document. Further it will be developing Dissemination and Handover document that will have three key constituents. The first describes the broad set of outputs, which is subject to update from time to time along with the mathematical logic behind the statistical model and the metrics created to understand performance of the GRO/mandal in liquor sales, as well as a clear documentation of the data. A second that explains the data process, from accessing the data in the API, to its processing, to the outputs created and where everything is subsequently stored. The third that is a compilation of fully commented statistical code to replicate all of the results, and which can be used to analyse any new data on liquor sales that becomes available for analysis. Taken together, these documents provide the

backbone for DAU to replicate all analysis, along with associated explanations. In the next phase, this analysis would be implemented and extended by staff of the Excise Department.

Iteration

A major aim of the DAU is to enable government departments to appreciate the value of processing administrative data to generate outputs that can support decision making. As the data becomes more "legible" to government departments, they will ask for adjustments and extensions to existing outputs, using their domain-specific knowledge to modify the statistical models developed by DAU. The final goal of such an iterative process would be to build capacity in the department to intelligently interpret data outputs, and eventually build and extend the models developed by the DAU.

Figure 6: Process Map: Iteration



We envision an iterative process in the following three steps:

1. **Understanding of Outputs:** At the first level, the DAU is entrusted with ensuring effective knowledge transfer to the user department and a helpful communication process, as outlined in the previous section, such that outputs are clearly communicated and understood. However, integrating and using the outputs for decision support will be a gradual process and the first level of iteration will involve refining the communication to maximize the benefits of analysis for decision support.

2. **Validity of Outputs:** The improvement in decision support is demonstrated only when the department's performance on key outcomes, e.g., rate of growth of tax revenue, or increasing the efficiency with which fraud is detected (i.e., maximizing the number of fraudulent taxpayers detected while minimizing the number of taxpayers incorrectly flagged as fraudulent), etc.
3. **Demand for New Outputs:** As departments become accustomed to analyzing data for decision support, and the initial outputs are validated, there will be demand for new types of outputs to, inter alia, improve departmental performance. They will also seek more control over the analysis of data, i.e., seek to build a departmental DAU.

Box 6: Example of Iteration

The current outputs from the statistical analysis in the Excise department has relied on daily sales of all types from a GRO. The focus has been on preventing leakage and ensuring that all sales are properly recorded and associated revenue realized. One natural extension of this approach is to try and extend it to different categories of alcohol, e.g., country liquor, IMFL, imported alcohol, etc. This can help to prevent stock-outs at particular GROs and help with inventory management. More into the future, to the extent possible, this can also build scenarios that tax at different rates, if the price response for each category is different.

HOW WILL DAU ACQUIRE CAPABILITIES?

The DAU will itself require support to build and strengthen its capabilities in data analysis. Its team is limited in size and will continue to require capabilities development support. This will be acquired through partnerships with reputed organisations and individuals with expertise in this area.

The Government of Andhra Pradesh has Memoranda of Understanding with reputed development institutions like the Development Monitoring and Evaluation Organisation (DMEO) located within the NITI Aayog, Centre for Policy Research (CPR), e-Government Foundation, and the Poverty Action Laboratory (J-PAL). These partnerships will be used to guide and develop capabilities of the DAU. The DAU is currently working closely with CPR to enhance its analytical capabilities on administrative data. DAU has also done deep dive workshops with e-Gov foundation to essentially develop a solutioning framework to adminstrative problem statements.

In addition, the Government has also constituted a Technical Advisory Group (TAG) of

reputed experts in the field to advise the Government and the DAU on data analytics-related issues. The TAG brings together experts from government, academia, think-tanks, technical domains, and private sector. They are assisting the DAU team with early direction, shaping expectations, and laying a foundation for a data-driven culture in the user departments.

WHAT WILL BE THE DAU STRUCTURE?

The DAU is a pioneering experiment within governments. It's therefore only natural that its scope of work, processes, nature of engagement, and structure will evolve over time.

While currently, there is a centralized DAU, with the head of DAU reporting to the Finance Department, specific teams will be developed for each departmental engagement. Ideally, these teams will comprise persons from DAU and persons from the department of engagement. While the persons from DAU will report into the head of DAU, the persons from the department of engagement will ideally report into an identified departmental nodal officer identified for this task. The nodal officer would be at a sufficient level to enable appropriate identification of what is needed by the department.

The departmental engagement team will brief the nodal officer on a regular basis, initially intensively, to ensure that the needs identification is done with clarity and consultation, since it forms the basis of deliverables for the various processes in the previous section. Subsequently, as the analysis proceeds, the DAU members will transfer skills to and build the capacity of the members from the department. Eventually, the members of the departmental team will form the nucleus of a customized department-level DAU. As it is formed, the customized departmental DAU will report into the departmental structure.

The centralized DAU will then move on to building capacity in another department and, for the departments with their own DAU, it will remain only as a facilitating unit for more complex analysis and implementation of frameworks that require coordination of data across multiple departments.

▪

DAU

DATA ANALYTICS UNIT (DAU) PROCESS DOCUMENT