CLEANSING PUNJAB SEWA-KENDRA DATA FOR IMPROVED SERVICE DELIVERY USING ANALYTICS, AI & BIG DATA – A CONCEPT NOTE

ABSTRACT

Existing Esewa records (> 3 Cr) related to Sewa Kendra transactions require cleaning in-order to determine actual number of unique citizens who may have availed of the services. Thereafter, the cleaned data can be used to create a Citizen Centric Data Model. This will help the State to build strategic dashboards, analyze scheme penetration and formulate better policies through transparency and cost efficiency.

Contents

1. Background

Punjab Sewa Kendras are a critical initiative by the Punjab Government aimed at providing efficient and accessible services to the citizens. These services span a wide range of areas, from welfare schemes to birth/death/marriage certificates. Sewa Kendras cater to 3.17 crores people of Punjab, with transactions amounting to approximately 8 lakhs per month [1]. However, one of the critical challenges the Sewa Kendra faces is the lack of a unique identifier for citizens. This absence makes it difficult to track and manage services provided to individual citizens effectively. Currently, citizens are identified primarily by their mobile numbers.

To address this, Punjab Govt. aims to develop a system that assigns a unique identification number to each citizen based on a combination of initials of the name, date of birth, district, and gender, ensuring persistence and uniqueness based on registration data. This would help create a citizen-centric view of services availed by each citizen over a period. This would help to achieve the following objectives:

- Enhanced service delivery with the ability to track and manage citizen interactions more effectively.
- Reduced duplication and misuse of services, resulting in cost savings for the Gov.

[1] [1] https://esewa.punjab.gov.in/

2. Approach

Each data project requires a unique approach to ensure its final dataset is reliable and accessible. The reliable dataset can then be leveraged to create a citizen centric services view, that would help manage any fraudulent benefits availed by the citizens, number of schemes they are enrolled in, and avoid overall duplication of benefits being availed.

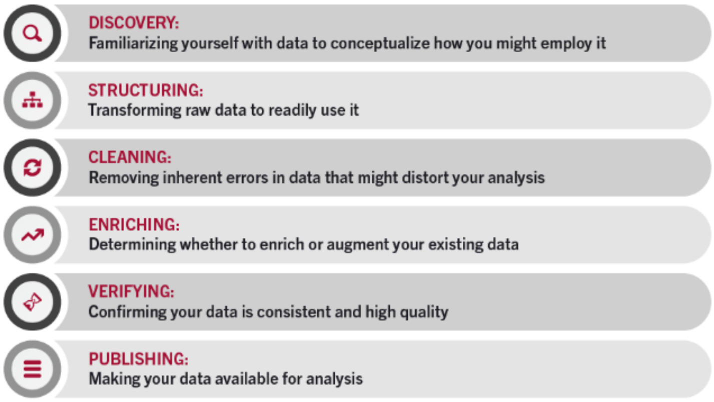
The objectives as stated above, can be achieved through the following sub-steps.

- ❖ Creation of a Unique 15-digit ID for every citizen who has availed or would avail services through Sewa Kendra (via the E-Sewa platform)
- ❖ Schemes, Citizen data mapping and Data Model
- ❖ Taluka, District wise services availed dashboards and Insights, etc.

Creation of Citizen Data-Record

The approach involves a series of steps known as data wrangling, which converts raw data into a more usable format. According to a Harvard Business School study2, there are six essential steps for creating the golden record or preparing data for any analytics project. These generic steps can be leveraged to generate distinct records and unique IDs for all new and existing Sewa Kendra users The steps are as follows:

Figure 1: Steps of Data Wrangling


DISCOVERY:
Familiarizing yourself with data to conceptualize how you might employ it
STRUCTURING:
Transforming raw data to readily use it
CLEANING:
Removing inherent errors in data that might distort your analysis
ENRICHING:
Determining whether to enrich or augment your existing data
VERIFYING:
Confirming your data is consistent and high quality
PUBLISHING:
Making your data available for analysis

1. Discovery

Discovery refers to the process of familiarizing with data so that one can conceptualize how one might use it. During discovery, one may identify trends or patterns in the data, along with obvious issues, such as missing or incomplete values, Duplicate records etc. that need to be addressed. This is an important step, as it will inform every activity that comes afterward.

2. Structuring

Raw data is typically unusable in its raw state because it's either incomplete or misformatted for its analytics. Data structuring is the process of taking this raw data and transforming it to readily usable for current and future analysis. The form the data takes would be Citizen Centric with a unique identifier for each citizen who has availed any of the services year to date.

3. Cleaning

Data cleaning is the process of removing inherent errors in data that might distort the analysis or render it less valuable. Cleaning can come in different forms, including deleting empty cells or rows,

[2] https://online.hbs.edu/blog/post/data-wrangling

removing outliers, merging records, removing duplicates, and standardizing inputs. The goal of data cleaning is to ensure there are no errors (or as few as possible) that could influence the final desired analysis. Identifying and removing any bad data greatly impacts the rest of the wrangling processes.

4. Enriching

Once one understands the existing data and has transformed it into a more usable state, one must determine whether one has all of the data necessary for the project at hand. If not, one may choose to enrich or augment the data by incorporating values from other datasets (Other departmental databases).

5. Validating

Data validation refers to the process of verifying that your data is both consistent and of a high enough quality. During validation, one may discover issues one needs to resolve or conclude through submitted documents or field surveys, etc. to have the data ready for analysis.

6. Publishing

Once the data has been validated, one can publish the Golden Records with the unique Citizen ID. This involves making it available to all departments within the Punjab.

Citizen-Centric Data Model

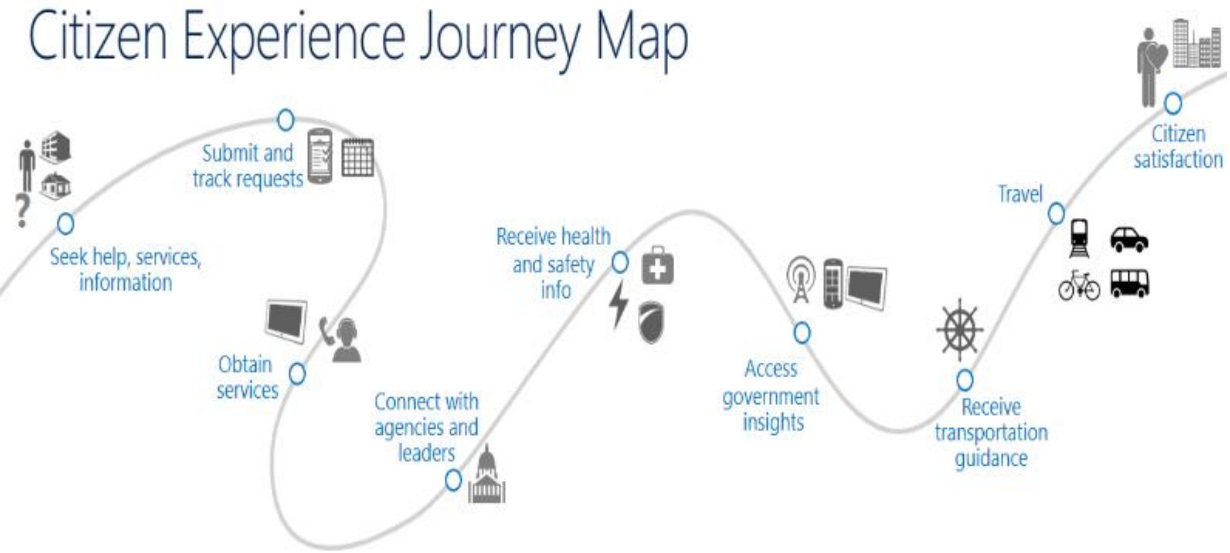
Delivering services to citizens is at the heart of what most government agencies do. In Punjab, the Sewa Kendra(s) are a key component in providing assisted service delivery to the citizens via the E-Sewa platform (used by the Sewa Kendra Operators).

E-Sewa is a critical platform that builds trust and shapes perceptions of the government. Citizens today expect transparent, accessible, and responsive services.

The traditional approach to designing and delivering government services has often centered around the government's own requirements and processes, rather than the needs of the people they serve. However, current efforts aim to shift this approach to a customer-centric model for service design and delivery through a "Citizen Journey".

A "citizen journey" is the entire experience that a person has when seeking a government service. The journey has a discrete beginning and end, and because it is typically multi-touch and multichannel, it is also cross-functional in nature. A citizen's journey is anchored in how people think about their experience, not in how government agencies do. This is best managed by pivoting the approach towards the citizen.

Figure 2: Citizen Experience Journey Map


Citizen Experience Journey Map
Seek help, services, information
Submit and track requests
Obtain services
Connect with agencies and leaders
Receive health and safety info
Access government insights
Receive transportation guidance
Travel
Citizen satisfaction

The key lies in building a unified data platform using a citizen-centric data model that describes and governs the data and is accessible to all personas in the ecosystem. The Citizen database (Master database) is to be kept at the center of the data model. The value of the data increases as we start enriching the data by leveraging other department's data sets or siloed data within the department.

For illustration, a Farmer-centric data model is shown in the attached figure. The data available on public portals for the farmers has been used to create the model. It is based on the following key entities:

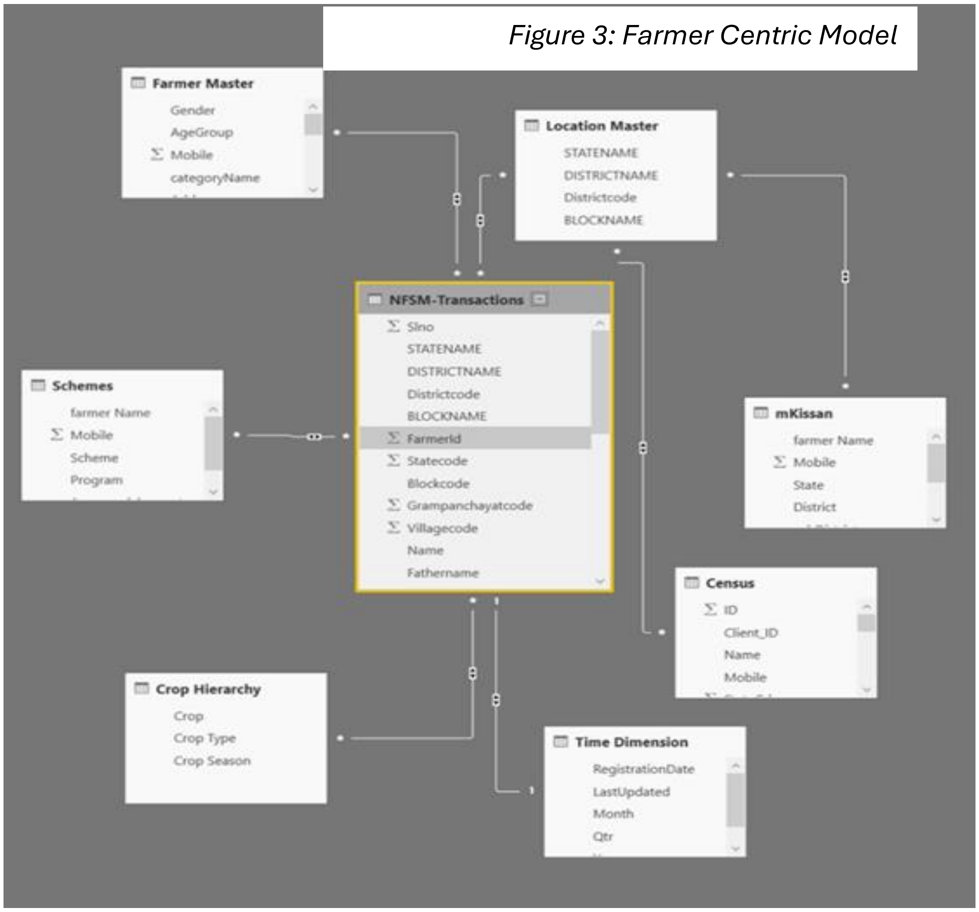


Figure 3: Farmer Centric Model

Individual (Farmer): To be derived using fuzzy matching and data quality processes, primarily leveraging Mobile Number, Name, Location details, with the latest update timestamp, which are common attributes in the provided data sets.

Location: To set up a standardized country hierarchy (based on census and state databases) - State -> District -> Block -> Village.

Crop: To be derived from the crop types based on the seasons and hierarchy thereof. For example, Rabi or Kharif and then Groundnut, Maize, Paddy, etc.

Time: To follow the described norm and would follow the time series - Months – Quarter – Year, mapped to crop season or financial calendar.

Measure: Land measurement unit. There are different unit measures across India and different meanings, e.g., Biswa in different states or within states have different measures, which would need to be optimized and standardized to create common units of measurement, e.g., Sq ft., Sq mt., Sq. Yard.

Scheme: Master table schemes would be required to manage and maintain various schemes, e.g., PMFBY, Soil Health Card.

Transaction Tables: All transactions related to schemes or specialized applications or subsidies based on department requirements would be independent transaction tables.

The data curated in the “Creation of Citizen data Record” shall be used at the center to map the citizen journey with the E-Sewa Platform in a similar approach."

Analytics & Dashboards

This involves the use of data analytics to improve the quality of public services. It is a way to measure the effectiveness of government programs and policies and to identify areas where improvements can be made. Citizen services analytics can help governments to better understand the needs of their citizens, and to provide more targeted and effective services.

The suggested approach to embark on finding insight, reducing duplication and misuse of services, and saving cost is the adoption of self-service analytics & intuitive tools, thus democratizing the power of data within an organization.

The dashboarding approach can be categorized as:

Operational dashboards: These dashboards monitor real-time data, track the end-to-end performance of various business operations, and communicate insights through clear data visualizations. They are designed to be used in daily workflows and provide time-sensitive insights and alerts about real-time developments. Operational E-Sewa Platform reports would fall under this category.

Figure 4: Dashboard for visualization

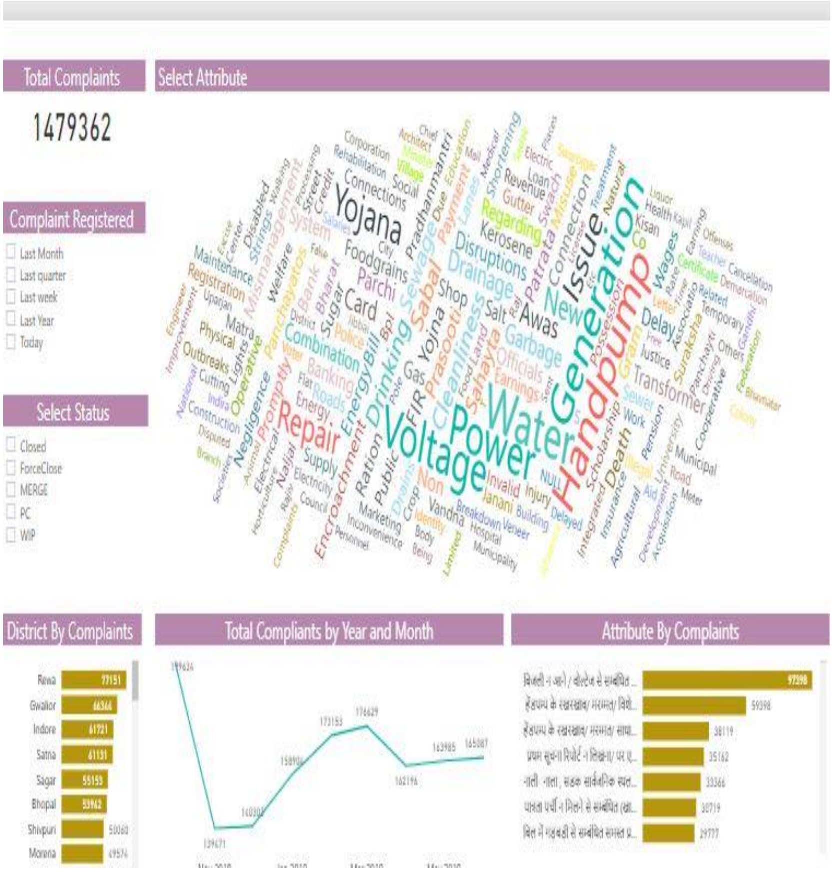
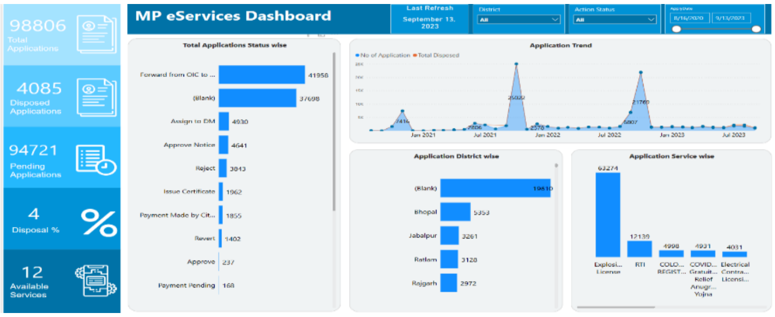
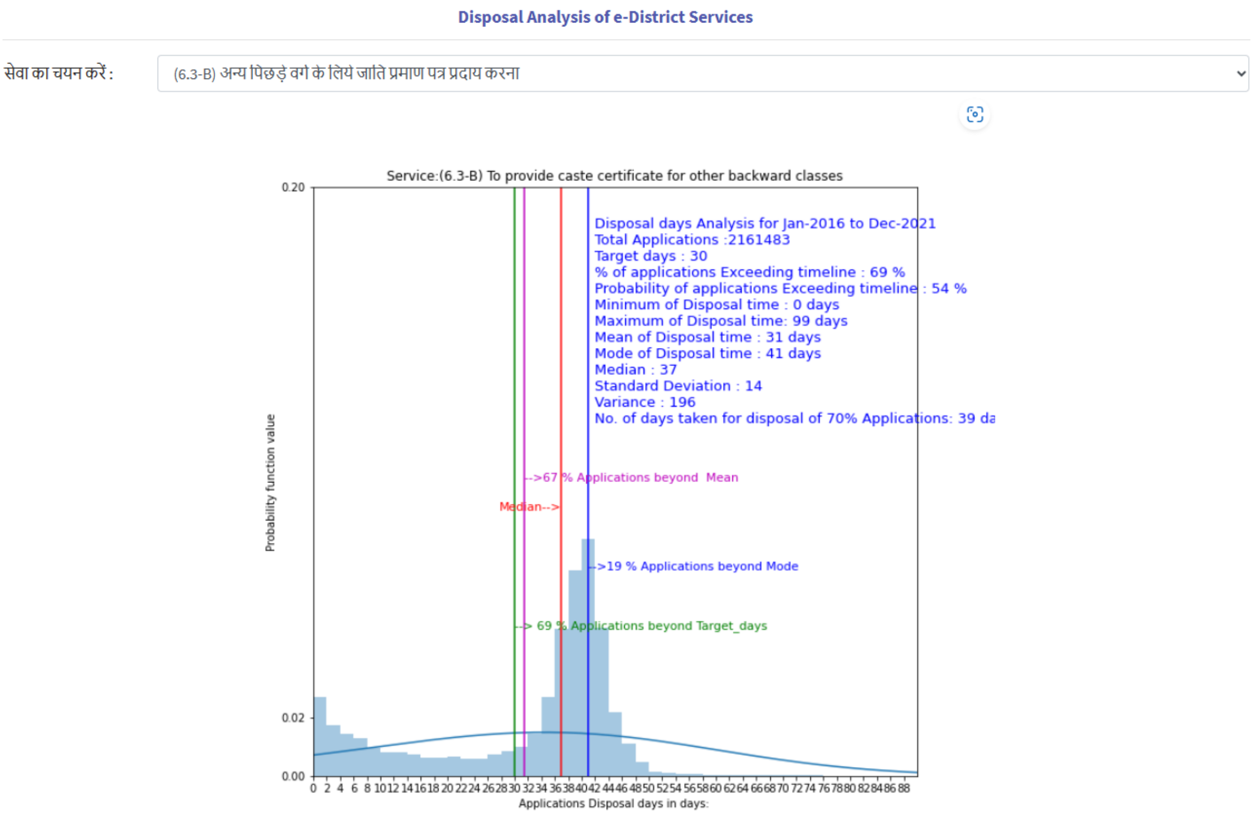


Figure 5: Strategic Dashboards



Strategic dashboards: These dashboards are mostly used by executives, directors, or business owners to monitor long-term KPIs. They provide high-level updates on the performance of the overall government programs. These dashboards update data less frequently than operational dashboards which are designed to be viewed daily.

Analytical dashboards: These dashboards are used to analyze and consume large amounts of complex data. Scheme wise benefits availed by Citizen of a household would fall in this category or Disposal Analytics as illustrated below.



Disposal Analysis of e-District Services

सेवा का चयन करें : (6.3-B) अन्य पिछड़े वर्ग के लिये जाति प्रमाण पत्र प्रदाय करना

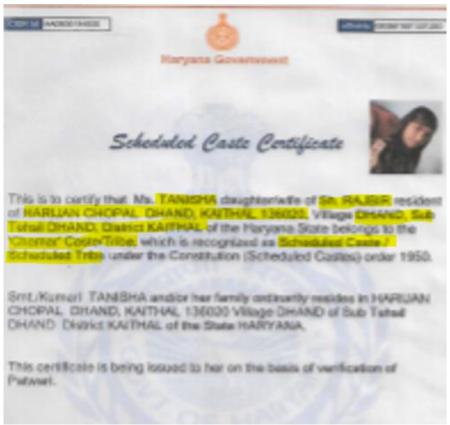
Service:(6.3-B) To provide caste certificate for other backward classes

Disposal days Analysis for Jan-2016 to Dec-2021
Total Applications :2161483
Target days : 30
% of applications Exceeding timeline : 69 %
Probability of applications Exceeding timeline : 54 %
Minimum of Disposal time : 0 days
Maximum of Disposal time: 99 days
Mean of Disposal time : 31 days
Mode of Disposal time : 41 days
Median : 37
Standard Deviation : 14
Variance : 196
No. of days taken for disposal of 70% Applications: 39 da

0.20
0.02
0.00
Probability function value
-->67 % Applications beyond Mean
Median-->
-->19 % Applications beyond Mode
--> 69 % Applications beyond Target_days
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84 86 88
Applications Disposal days in days:

Leverage **Artificial Intelligence** based ready to use Optical Character Recognition pipeline (for custom forms such as cast certificate, school marksheets, school leaving certificate, etc.) to create an algorithm to precisely crop out the region of interest, e.g., Name, date of birth, etc. details form certificate submitted, & match it with the submitted forms to bring process efficiencies by reducing the time and manual efforts for validation of the data.

Let's take one sample document, cast certificate. The cast certificate (sample) embedded below for illustration. The certificate contains the details on the name of the citizen, father / husband's name, 'resident of' and 'caste' and classification as SC / ST, Backward Class, etc.



CIDR Id HA0900184006 [illegible]

Haryana Government

Scheduled Caste Certificate

This is to certify that Ms. TANISHA daughter/wife of Sh. RAJBIR resident of HARIJAN CHOPAL DHAND, KAITHAL 136020, Village DHAND, Sub Tehsil DHAND, District KAITHAL of the Haryana State belongs to the 'Chamar' Caste/Tribe, which is recognized as Scheduled Caste / Scheduled Tribe under the Constitution (Scheduled Castes) order 1950.

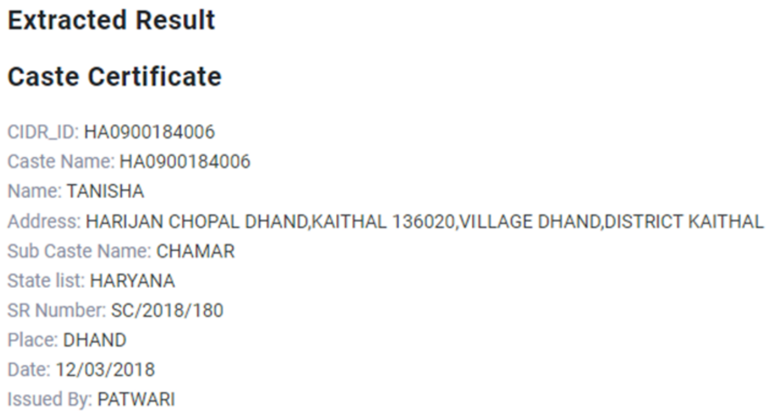
Smt./Kumari TANISHA and/or her family ordinarily resides in HARIJAN CHOPAL DHAND, KAITHAL 136020 Village DHAND of Sub Tehsil DHAND District KAITHAL of the State HARYANA.

This certificate is being issued to her on the basis of verification of Patwari.

Name	Father's Name
Address	Village
Tehsil	District
Cast	Classification

Verified By

The relevant details as highlighted above are extracted using the form recognizer custom key value pairs using automated algorithms or artificial intelligence.



Extracted Result

Caste Certificate

CIDR_ID: HA0900184006
Caste Name: HA0900184006
Name: TANISHA
Address: HARIJAN CHOPAL DHAND,KAITHAL 136020,VILLAGE DHAND,DISTRICT KAITHAL
Sub Caste Name: CHAMAR
State list: HARYANA
SR Number: SC/2018/180
Place: DHAND
Date: 12/03/2018
Issued By: PATWARI

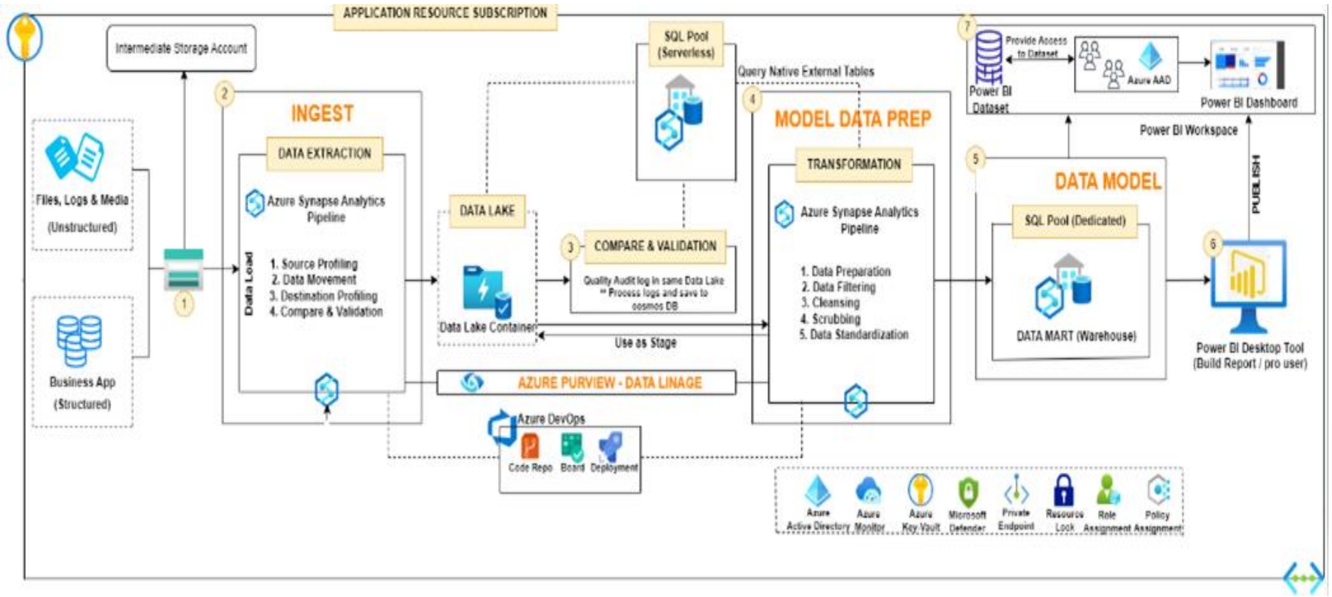
The proposed solution extracts the fields from the cast certificate and makes it available for cross reference with the filled values in the family database against member details CastCategory, FirstName, LastName, District, Tehsil (Address), etc. column. Each extraction is accompanied by the model confidence score. The scrutinizing authority can choose to accept the verification results over certain confidence stores and pass on the rest for manual verification.

In addition, according to a report by Harvard Ash Center, artificial intelligence (AI) can be used to improve citizen services in government. AI can help reduce administrative burdens, resolve resource allocation problems, and take on complex tasks. Many AI case studies in citizen services today fall into five categories answering questions, filling out and searching documents, routing requests, translation, and drafting documents. [3]These applications could make government work more efficient while freeing up time for employees to build better relationships with citizens. With citizen satisfaction with digital government offerings leaving much to be desired, AI may be one way to bridge the gap while improving citizen engagement and service delivery

[3] https://ash.harvard.edu/files/ash/files/artificial_intelligence_for_citizen_services.pdf

3.Architecture & Flow

[4]The section highlights the high-level architecture with details of the services that shall be leveraged in order meet the project objectives through the approaches described above.



1. **Data Ingestion:**

   Relational Databases: Use Azure Synapse pipelines to pull data from a wide variety of databases, both on-premises and in the cloud. Pipelines can be triggered based on a pre-defined schedule, in response to an event, or can be explicitly called via REST APIs.

   Semi-Structured Data: Use Azure Synapse pipelines to pull data from a wide variety of semi-structured data sources, both on-premises and in the cloud. For example: Ingest data from file-based sources containing CSV or JSON files or Connect to No-SQL databases such as Cosmos DB or Mongo DB or Call REST APIs provided by SaaS applications that will function as your data source for the pipeline.

   Non-Structured Data: Use Azure Synapse pipelines to pull data from a wide variety of non-structured data sources, both on premises and in the cloud. For example: Ingest video, image, audio, or free text from file-based sources containing the source files or Call REST APIs provided by SaaS applications that will function as your data source for the pipeline.

2. **Data Storage & Transformation (Cleansing)**

   From the Azure Synapse pipeline, one can use a Copy Data activity to stage the data copied from the different sources into the Raw zone of Azure Data Lake Store Gen 2 data lake. One can save the data in delimited text format or compressed as Parquet files. Leverage Data Flows, SQL Serverless queries, or Spark notebooks to validate, transform, and move the datasets into

[4] The architecture and services suggested in this section are based on Azure Services from Microsoft as it is the current cloud provider for DoGR (Sept 2023)

Curated zone in data lake. For semi-structured data SQL Serverless queries expose underlying CSV, Parquet or JSON files as external tables so they can be queried using T-SQL. Unstructured data can be saved preserving the original format as acquired from the data sources.

Apart of your data transformations, one can invoke machine learning models from your SQL pools using standard T-SQL or Spark notebooks. These ML models can be used to enrich datasets and generate further business insights. These machine learning models can be consumed from Azure Cognitive Services or custom ML models from Azure ML.

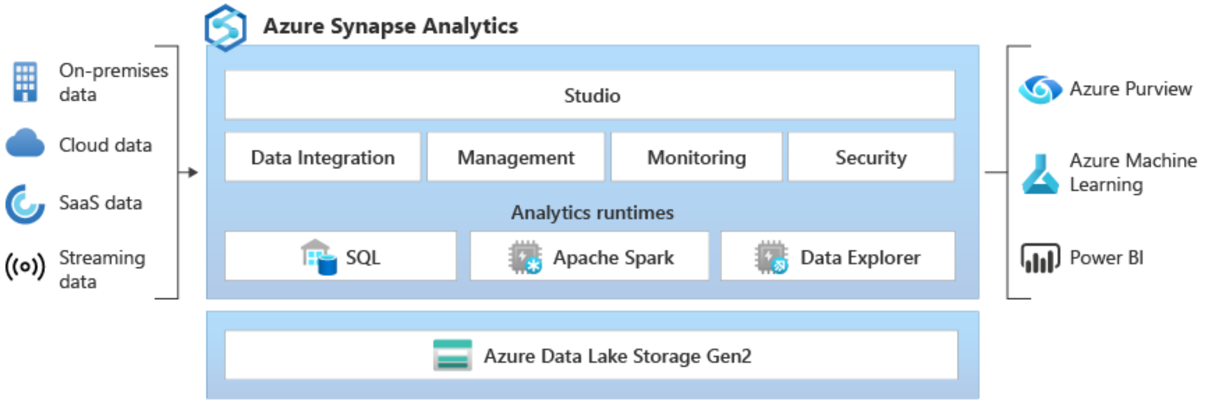
3. **Data Models & Analytics**

The final dataset can be served directly from the data lake Curated zone or by using Copy Data activity to ingest the final dataset into SQL pool tables using the COPY command for fast ingestion.

Load relevant data from the Azure Synapse SQL pool or data lake into Power BI datasets for data visualization. Power BI models implement a semantic model to simplify the analysis of business data and relationships. Business analysts use Power BI reports and dashboards to analyze data and derive business insights.

4.Azure Services

1. **Azure Synapse** is an **enterprise analytics service** that accelerates time to insight across data warehouses and big data systems. Azure Synapse brings together the best SQL technologies used in **enterprise data warehousing**, Spark technologies used for big data, Data Explorer for log and time series analytics, Pipelines for data integration and ETL/ELT, and deep integration with other Azure services such as Power BI, Cosmos DB, and AzureML



Synapse SQL uses a scale-out architecture to distribute computational processing of data across multiple nodes. Compute are separated from storage, which enables you to scale compute independently of the data in your system. For dedicated SQL pool, the unit of scale is an abstraction of compute power that is known as a data warehouse unit.

For serverless SQL pool, being serverless, scaling is done automatically to accommodate query resource requirements. As topology changes over time by adding, removing nodes or failovers, it adapts to changes and makes sure your query has enough resources and finishes successfully. For example, the following image shows serverless SQL pool using four compute nodes to execute a query. What is Azure Synapse Analytics? - Azure Synapse Analytics | Microsoft Learn

2. **Apache Spark** is a powerful data processing engine for Big Data analytics. Spark processes data in small batches and is a parallel processing framework that supports in-memory processing to boost the performance of big-data analytic applications. Apache Spark currently supports Python, R, and Scala. PySpark is a python flavor of Apache Spark. Apache Spark Dataframes provide a powerful and flexible toolset for cleaning and preprocessing data.

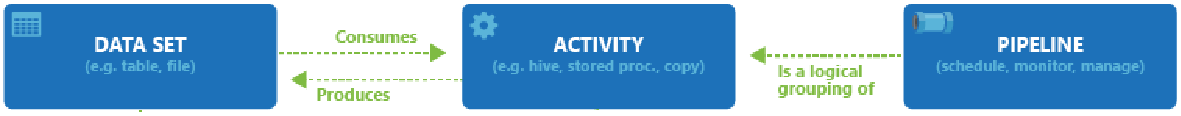
   Apache Spark in Azure Synapse Analytics is one of Microsoft's implementations of Apache Spark in the cloud. A serverless Apache Spark pool, when instantiated, is used to create a Spark instance that processes data. When a Spark pool is created, it exists only as metadata, and no resources are consumed, running, or charged for. A Spark pool has a series of properties that control the characteristics of a Spark instance. These characteristics include but aren't limited to name, size, scaling behavior, time to live. Spark instances are created when one connects to a Spark pool, creates a session, and runs a job. When one submits a second job, if there's capacity in the pool, the existing Spark instance also has capacity. Then, the existing instance will process the job. Otherwise, if capacity is available at the pool level, then a new Spark instance will be created. Azure Synapse makes it easy to create and configure Spark capabilities in Azure. Apache Spark in Azure Synapse Analytics overview - Azure Synapse Analytics | Microsoft Learn

3. **Data lake** is a single, centralized repository where one can store all data, both structured and unstructured. A data lake enables any organization to quickly and more easily store, access, and analyze a wide variety of data in a single location. With a data lake, one doesn’t need to conform the data to fit an existing structure. Instead, one can store the data in its raw or native format, usually as files or as binary large objects (blobs).

   Azure Data Lake Storage is a cloud-based, enterprise data lake solution. It's engineered to store massive amounts of data in any format, and to facilitate big data analytical workloads. One can use it to capture data of any type and ingestion speed in a single location for easy access and analysis using various frameworks. Azure Data Lake Storage Gen2 Introduction - Azure Storage | Microsoft Learn

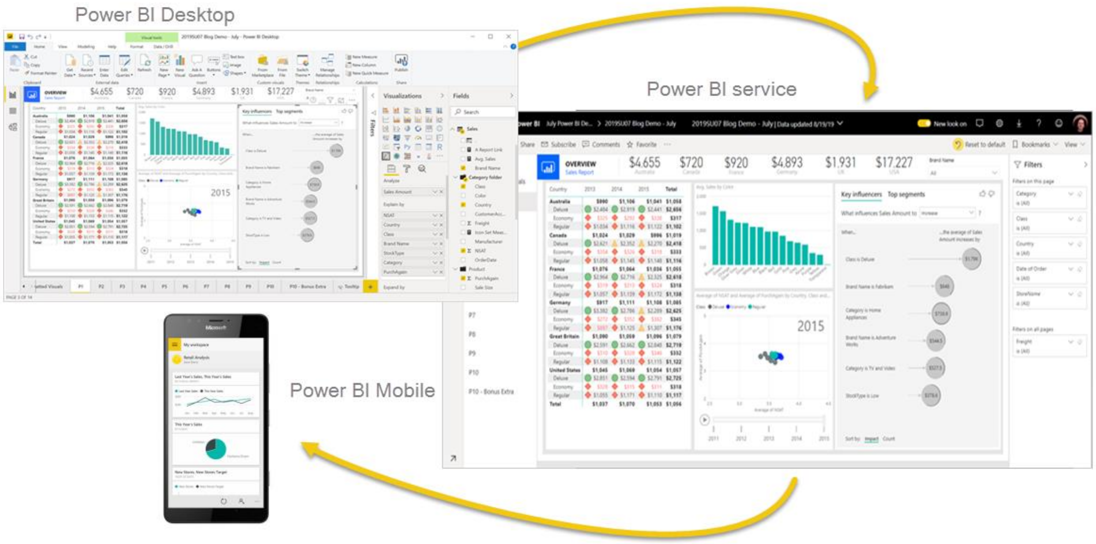
4. **Azure Synapse Pipelines** is a service that allows one to create data-driven workflows for orchestrating data movement and data processing at scale. One can author and monitor pipelines using Synapse Studio. Pipelines can integrate with other Azure services such as SQL, Data Explorer, Power BI, CosmosDB, and AzureML. Pipelines can also support big data compute services such as HDInsight and DataBricks.

   Azure Synapse Analytics pipelines have three groupings of activities: data movement activities, data transformation activities, and control activities. An activity can take zero or more input datasets and produce one or more output datasets. The following diagram shows the relationship between pipeline, activity, and dataset:



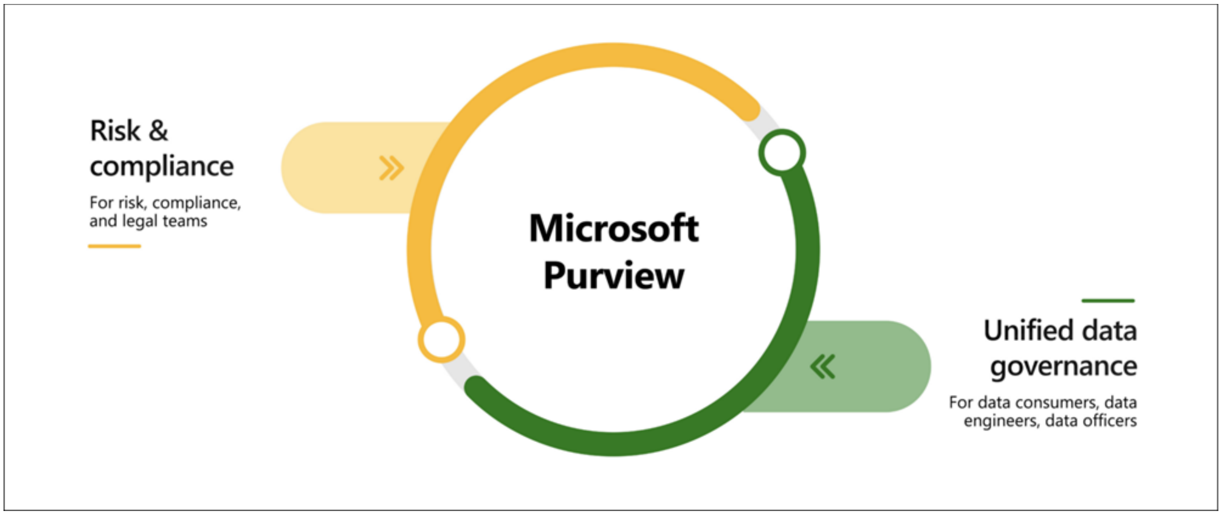
A pipeline is a logical grouping of activities that together perform a task. For example, a pipeline could contain a set of activities that ingest and clean data, and then kick off a mapping data flow to analyze the data. The pipeline allows one to manage the activities as a set instead of each one individually. One deploys and schedules the pipeline instead of the activities independently. [Pipelines and activities - Azure Data Factory & Azure Synapse | Microsoft Learn]

5. **Power BI** is a collection of software services, apps, and connectors that work together to turn your unrelated sources of data into coherent, visually immersive, and interactive insights. Your data might be an Excel spreadsheet, or a collection of cloud-based and on-premises hybrid data warehouses. Power BI lets you easily connect to your data sources, visualize and discover what's important, and share that with anyone or everyone you want.



Power BI datasets are used for data visualization. Power BI models implement a semantic model to simplify the analysis of business data and relationships. Business analysts use Power BI reports and dashboards to analyze data and derive business insights. [What is Power BI? - Power BI | Microsoft Learn]

6. **Microsoft Purview** is a family of **data governance, risk, and compliance** solutions that can help your organization govern, protect, and manage your entire data estate. Microsoft Purview solutions provide integrated coverage and help address the recent increases in remote user connectivity, the fragmentation of data across organizations, and the blurring of traditional IT management roles. [What is Microsoft Purview? | Microsoft Learn]


Risk & compliance
For risk, compliance, and legal teams
Microsoft Purview
Unified data governance
For data consumers, data engineers, data officers

- Use Azure Purview for data discovery and governance insights on your data assets, data classification and sensitivity covering the entire organizational data landscape.
- Azure Purview can help you maintain a business glossary with the specific business terminology required for users to understand the semantics of what datasets mean and how they are meant to be used across the organization.
- You can register all your data sources and setup regular scans to automatically catalog and update relevant metadata about data assets in the organization. Azure Purview can also automatically add data lineage information based on information from Azure Data Factory or Azure Synapse pipelines.
- Data Classification and Data Sensitivity labels can be added automatically to your data assets based on pre-configured or customs rules applied during the regular scans.
- Data governance professionals can use the reports and insights generated by Azure Purview to keep control over the entire data landscape and protect the organization against any security and privacy issues.