

## Topic 11

1. **Title: *Intelligent Data Catalogue***
2. **Description:** Bhuvan is a host of multitude of datasets generated from a wide variety of sources spanned across different services in the platform. With this large amount of data and its diversity, it is in Bhuvan, an efficient approach to develop an intelligent data cataloguing service shall be developed which will be helpful for quick querying for the user. A data catalogue system helps in organizing the data, aids in quick search and access, to make analytics and meaningful insights from the data.
3. **Objectives:**
  - a. To build an organized inventory of data assets and develop a data catalogue.
  - b. Automatically tag and classify the incoming data and update the catalog on realtime using automation techniques for better data governance.
  - c. To provide efficient data search mechanism and discovery for users.
  - d. Convert the data catalogue service as an API to be integrated with other services.
4. **Expected Outcomes:**
  - a. Data cleansing and classification of the incoming datasets.
  - b. A structured data catalog system along with a search feature.
  - c. A data catalogue service that can accessed as an API for integration.
5. **Relevant data and steps to get the data from Bhuvan/ other sources:**

Test Datasets are provided in the repository (link details in the web page)
6. **Steps to be followed for achieving the objectives:**
  - a. Develop an ML model for data classification of the input data.
  - b. Use the free datasets to train the model for data classification.
  - c. Create a sample catalogue system and populate the free datasets/ input datasets.
  - d. Integrate a search and query tool.
  - e. Automate the process to clean, classify and organize any type of incoming data.
  - f. Bundle and develop an API to consume this as a service for integration with other tools/services.
7. **Evaluation**
  - a. Algorithm will be run to build Data Catalogue for sample datasets.
  - b. Search feature will be tested to identify the data from the catalogue displaying the location and metadata details of the file/dataset.
  - c. API will be Integrated with Bhuvan platform and verify for sample project datasets.