Data Science tools - Process Documentation

Meta description/Summary:

A myriad of data science tools and techniques are used by the Abhyaz teams on the Abhyaz platform. Data science applications mostly include data collection and analysis processes. Data science skills are quite high in demand across all industries now. This is mainly due to its capacity to provide accurate insights and projections based on quantitative data. *The utilization of data science tools in Abhyaz has been summarized in this paper along with a description of how these deliver effectiveness*.

Definition of the process:

Abhyaz teams use multiple types of data science tools to provide hands-on experience to our platform users and interns. We use tools such as ***Octoparse, Webscrapper.io, Skrapp.io, Apollo.io, Email extractor, and Google Dorking*** to collect and process data. New-age data analytics techniques are efficiently used to identify market requirements and skills in demand. Teams collect large amounts of data from different sources. These provide insightful details about the internal performance of teams. Data science teams are also directly involved in analyzing external market conditions in relevant industries.

Process scope

Implementation of data science technologies has become quite popular in improving learning efficiency across industries. Data science tools are used on Abhyaz platforms for using data to leverage outcomes in training. *The main purpose of using data science tools in Abhyaz is to achieve accuracy in knowledge mapping, predictive analyses and design programs that cater to variable individual learning needs.*

Purpose

The data science tools used on the Abhyaz teams are mostly used for supporting the ***ATP and ASP programs*** provided by Abhyaz. The purposes of data collection and use of different data science in Abhyaz are:

- *To carry out industrial research*
- *To market the products to target customers segments,*
- *To provide freshers with industrial experience opportunities in data science fields.*

Abbreviations/Acronyms

- **ATP:** Abhyaz Talent Program
- **ASP:** Abhyaz Skill Program
- **LMS:** Learning Management Systems
- **DA:** Data Analytics
- **TPOs:** Training and placement officers
- **AVW:** Abhyaz Virtual Workplace
- **CRM:** Customer Relationship management
- **OSINT:** Open source intelligence

Procedures

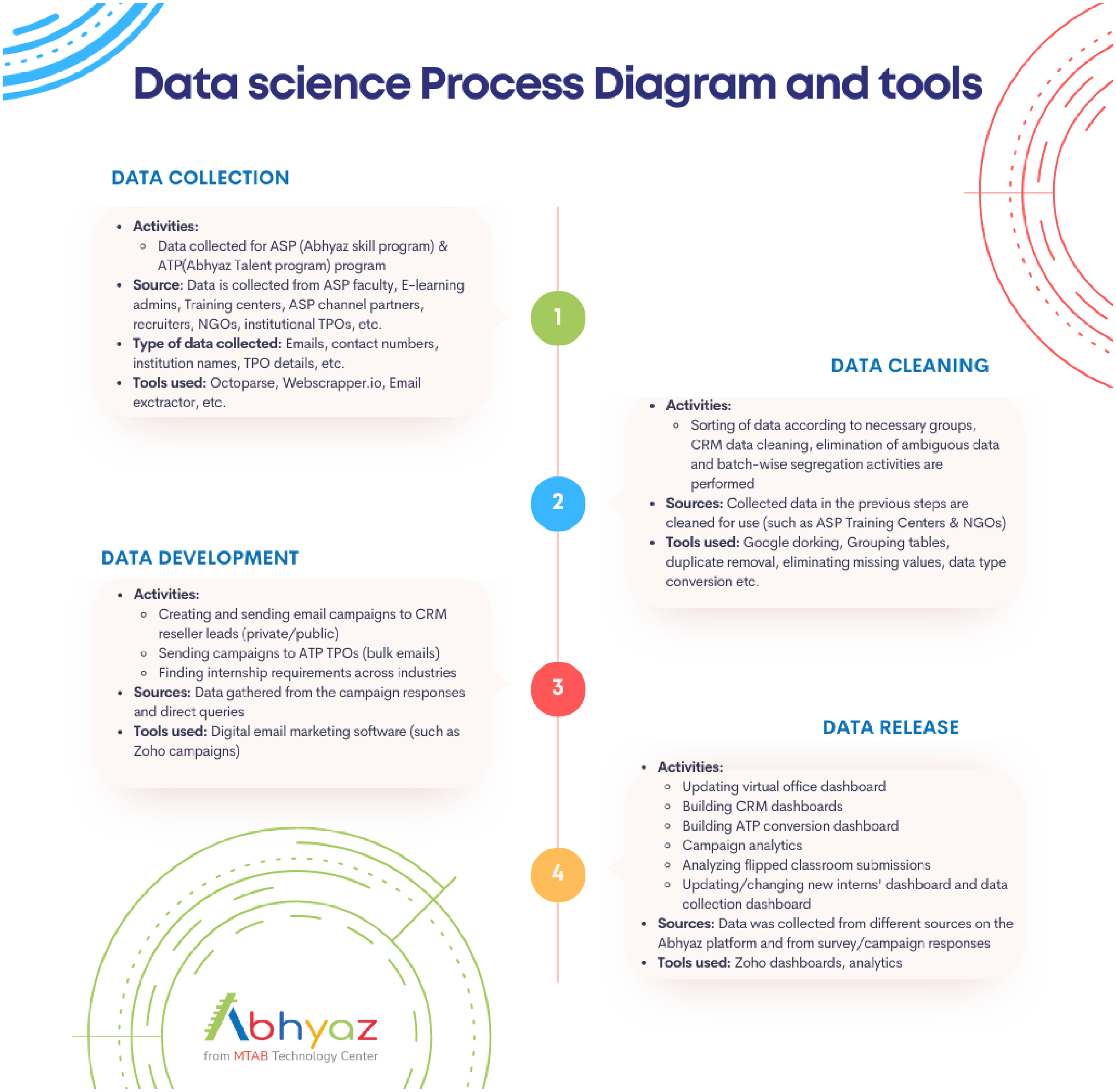
The current process of using different data science tools in Abhyaz can be divided into a few specific steps. These are:

- ***Step 1. Data collection:*** Firstly, our data science team collects large amounts of data from different sources online for analysis. In this step, the primary data types include ***emails*** and ***contact details*** of institutional personnel. Both the ATP and ASP programs require different sets of relevant data. In ATP, internships are provided to students. Therefore, collected data include ***contact emails and phone numbers*** of ***TPOs*** in recruiting institutions. In ASP, on the other hand, interns are exposed to other companies. Hence, data related to HR recruitment processes and skill requirements are collected. In this step, the data science tools such as ***Octoparse, Web scrapper, Email extractor, Google Dorking*** etc. are used by the data science team. In some cases, our data science teams also administer online forms, bulk email campaigns and questionnaires for collecting data.
- ***Step 2. Data cleaning:*** The second step in the data analysis process is the data cleaning step. This step involves ***sorting, organizing, updating or eliminating*** data in the data

sets. We categorize and group the data collected according to necessities. For example, all public and private domain emails are separately grouped. In this step, we also eliminate incomplete data and redundant data. ***Batch-wise segregation*** is done in this step. Data science functions that are useful in this stage are duplicate elimination or replacement, data type conversion tools and others. Internal teams use application functions in ***MS Excel, Google Sheets or Zoho sheets*** for this at Abhyaz.

- ***Step 3. Data development:*** The third step is the Data development phase. After cleaning the collected data, we send bulk emails to different groups created from the data sets. ***ATP TPOs*** and ***CRM reseller leads*** are the recipients of such email campaigns. Responses to these campaigns are also collected for future analysis. ***Zoho campaigns, forms and Zoho surveys*** are used to design the response sheets and questionnaires at Abhyaz. Through this, we can gather useful market insights about skill demands in various technical fields and recruitment patterns.
- ***Step 4. Data release:*** The Abhyaz team activities in this step include building analytics dashboards using tools such as ***Zoho analytics***. Data collected from campaigns, interns' performance data as well as conversion rate data are the basic categories. The platform is also capable of automatically showing visualizations of the collected data using ***Zoho dashboards***. Associated activities in this step include ***data elaboration*** and ***data testing*** as well.

Process Flow diagram



Data science Process Diagram and tools

DATA COLLECTION

1

- **Activities:**
  - Data collected for ASP (Abhyaz skill program) & ATP(Abhyaz Talent program) program
- **Source:** Data is collected from ASP faculty, E-learning admins, Training centers, ASP channel partners, recruiters, NGOs, institutional TPOs, etc.
- **Type of data collected:** Emails, contact numbers, institution names, TPO details, etc.
- **Tools used:** Octoparse, Webscrapper.io, Email exctractor, etc.

DATA CLEANING

2

- **Activities:**
  - Sorting of data according to necessary groups, CRM data cleaning, elimination of ambiguous data and batch-wise segregation activities are performed
- **Sources:** Collected data in the previous steps are cleaned for use (such as ASP Training Centers & NGOs)
- **Tools used:** Google dorking, Grouping tables, duplicate removal, eliminating missing values, data type conversion etc.

DATA DEVELOPMENT

3

- **Activities:**
  - Creating and sending email campaigns to CRM reseller leads (private/public)
  - Sending campaigns to ATP TPOs (bulk emails)
  - Finding internship requirements across industries
- **Sources:** Data gathered from the campaign responses and direct queries
- **Tools used:** Digital email marketing software (such as Zoho campaigns)

DATA RELEASE

4

- **Activities:**
  - Updating virtual office dashboard
  - Building CRM dashboards
  - Building ATP conversion dashboard
  - Campaign analytics
  - Analyzing flipped classroom submissions
  - Updating/changing new interns' dashboard and data collection dashboard
- **Sources:** Data was collected from different sources on the Abhyaz platform and from survey/campaign responses
- **Tools used:** Zoho dashboards, analytics

Abhyaz
from MTAB Technology Center

Figure 1: Process flow diagram and tools used in Data science

Identification of involved parties

In each step of our data science application process, various partners and providers are involved:

- ***Step 1:*** ASP Faculty, E-Learning Administrators, ASP Training Centers, NGOs, Indian or Non-Indian ASP Channel Partners, ATP TPOs, recruiters and Industrial partners.
- ***Step 2:*** CRM Leads, Contacts & Accounts, Faculty and administrators, Data analysts.
- ***Step 3:*** CRM reseller leads, ATP TPOs, Data analysts, DA interns
- ***Step 4:*** ATP TPOs, Data analysts, DA interns, Faculty and administrators

Exceptions and control points

Some exceptions that potentially occur in the application of data science tools can be listed as follows:

Exceptions	Control points
Inappropriate or unnecessary data collection	➢ Data-sheets gathered from the responses are validated and reviewed by E-Learning Administrators and internal DA team leads
Data science teams may face issues such as missing data in some particular responses. This mainly happens when respondents leave certain fields blank while responding to surveys.	➢ Columns with missing values can be dropped if they contain irrelevant data. ➢ Alternatively, missing values are replaced by mode/median values. ➢ The most frequent value can be used and interpolation techniques are also applied sometimes.
Similar data rows can be repeated in the data set	➢ Deduplication techniques are used by the DA teams ➢ Columns are sorted using application functions in Zoho sheets or excel sheets

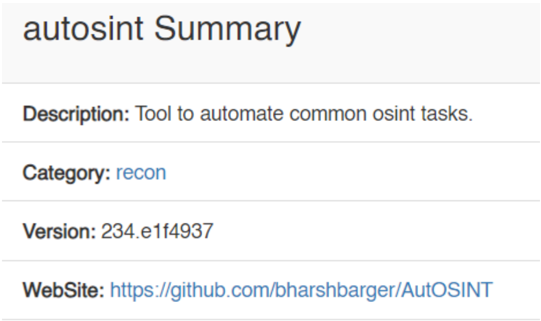
Resources

- We collect data from ASP Channel Partners (Indian + Non-Indian).
- Intern onboarding and performance data is also an important source.
- ATP Conversion data, Flipped classroom submissions data, as well as Campaign response data, are also used for Data analytics at Abhyaz.
- Other sources include ATP Recruiters Data, data from ASP Training Centers & NGOs

Future state

The Data Science teams in Abhyaz have planned to further improve the data analytics outcomes by incorporating new tools into the processes. Some of the tools that are to be used in data science include:

- **AutOSINT:** This is a tool for automating OSINT tasks. this is one of the tools used for penetration testing.
- **theHarvester**: This tool is also used in data science activities such as penetration test or red team engagement. This tools is especially helpful for gathering OSINT -data that helps in determining the external threat landscape of organizations online.



autosint Summary

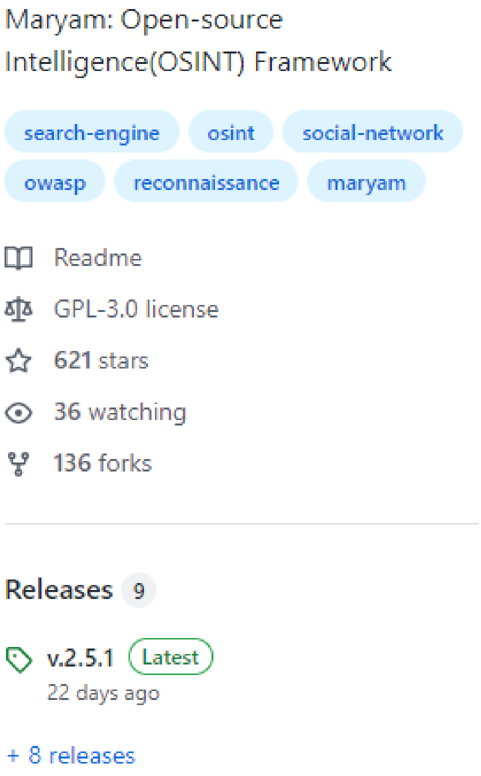
Description: Tool to automate common osint tasks.

Category: recon

Version: 234.e1f4937

WebSite: https://github.com/bharshbarger/AutOSINT

- **OWASP Maryam**: This tool is an optional or modular open source OSINT-based framework which is used for data gathering. It is written in the Python language and it provides a powerful data-harvesting environment from open sources.
- **Infoga - Email OSINT**: This is another tool useful for information extraction specifically from public source email accounts (Search engines, pgp key servers or shodan).
- **Photon:** This is a crawler tool designed for OSINT. It can extract data such as URLs, Files, Intel, Secret keys, sub-domain and DNS data, etc.
- **Plessas Expert Network:** This tool "*provides effective analysis with proven intelligence techniques and methodologies*".



Maryam: Open-source Intelligence(OSINT) Framework

All of these tools are mostly useful in the data collection stage. Some data science tools like tableau, Python/R programming, Rapidminer, etc. have become indispensable at present. These data analysis tools are also to be used for analysis and visualization of data in future.