**Infrastructure**

1. ELK stack (Big Data Processing and Analytics) - ElasticSearch, Logstash, Kibana.
2. Programming Languages – Java 8,9, Python, GoLang
3. Spring-boot, Hibernate
4. Web Crawling at a very large scale – Scrapy, Image/banner crawler for automated hyper personalisation
5. Machine Learning/Deep Learning, NLP
6. Sentence Transformers, Keras, TensorFlow, Theano, Gensim, Numpy, Scipy, sklearn, XGBoost, NLTK, Word Embedding(word2vec), Glove, PySparNN for sparse data, Annoy, Text rank, DBpedia, SPARQL
7. Concurrency frameworks – Executor Service, Cognite Process Pool
8. GraphVite, Pykeen, Fast text, KG2Vec

Knowledge graph AugmentedNeuralNetwork, Embedding matrix derived from embeddings from different databases (wikipedia, twitter, conceptnet) (vectors from word2vec, sentence transformers)

1. Ethereum layer 2 chain framework, transfer gateways, allegro big cache, DApps, cryptocurrency/wallet design, online rewards management, trading card game, interoperable blockchains (transfer gateways), Go based smart contracts, Solidity based smart contracts, decentralized chat server application, Adserver application, social network application, smart contract based ad fraud prevention,(impression, click event trackers storage in blockchain via Go based smart contract transactions (scaled using layer 2 chains), Block Explorer, Block/Transaction Level DB database, RPC API layer for data fetch from blockchain, DPOS consensus engine, coin minting policies, Whitelisting smart contracts). Tendermint based - Full implementations available. Blockchain networks used – Ethereum, Tron, Binance.

**Key value stores**

1)**LevelDB**

2)**BadgerDB** ( ~4 times faster Get and a potential 86% reduction in AWS bills, due to less reliance on RAM and more reliance on ever faster and cheaper SSDs)

3)**BoltDB**

For vector storage (bloom filter, lru cache implementations in level db are also used) and other applications.

All user vectors are serialised to level db on hourly basis for daily Aajtak/IndiaToday traffic.

This leveldb stores time series graph of user vectors for full Aajtak/IndiaToday Traffic.

Applications of user vectors log is -

1)Using bloom filters for fast log scanning

2)

i)To check presence of a user in particular month, this data is used by algorithm which determines loyalty nature (useful for deriving churning nature of users given a semantic context (set of topics/data points))

ii)To efficiently derive semantic fingerprint of a user at a particular point in time, by using bloom filters. (User vectors can be compared with semantic context vectors (set of topics/data points))

1. Caching Frameworks – Ehcache (Distributed Cache, Consistent Hashing based Port available for scaling purposes), Allegro big cache, Go based Arc, Two Queue cache.

Applications –

1)Used for Query Cache (First Publisher Dashboard to reduce API latencies)

2)Used for Persona Cache (Persona Warehouse for firstpartyId based Personas)

Persona WareHouse Memory is optimised by using first party Id relevant to current campaign schedules, first party Ids which are loyal/returning in nature

Ehcache uses Direct Memory for storing bulk of Personas.

1. Data Structures - Suited for Big Data

1)External Sort Algorithm

A multi-way external merge sort that use a TreeMap (Java Red-Black Tree) as

-internal sorting data structure

-is able to scale over GBs of data better than the unix sort command.

2)Fast buffered writer

1. Front End
2. Angular JS, JQuery
3. Lighthouse profiling scripts
4. Instrumentation tools (API Instrumentation) - Prometheus and Grafana
5. E2E Tests
6. Build Management tools - Travis and Jenkins (e2e test reports verification for new builds)

14. Slack Integration for alerts/notifications, pager duty.

15. New Relic/Zabbix for server monitoring/ES cluster monitoring.

16. Logging Server for System Reports (Load Balancer, Ingestion Server, Crawler, Enhancer Modules, Persona Derive, Persona Sync, ML signals)

## **Modules Brief Description**

## **Tag Manager**

* It includes Implementation of Tag Manager for data collection from different sources such as PCs, mobile web, Game Consoles, Smart TVs.
* It tracks different events such as Shopping cart oriented events, items added, checkout stage, purchase stage.
* Collects more than 100 data points of users.
* Android/iOS based SDK for collecting data points from Apps and streaming data to datawarehouse. (lightweight, Disk cache, Offline mode (to prevent data loss), Wi-fi only mode, Bulk Urls processor for GAIDS/IDFA)
* Data points collected from different channel types such as PCs, mobile web, Apps are mapped to users to get a unified 360 degree profile of user.
* User of Beacons.
* Id Management – First party Id, Session based Id, Third Party Id and Corresponding Maps.
* Cookie Syncing code with other Data providers and DMPs/CDPs.
* Persona Sync javascript (async persona synchronization in local storage using async await, on the fly topic targeting support present in urls), (plug and play, enable and disable) personas different domain personas, persona cleaning support.

## **Semantic Engine**

Solutions composed of -

1. **Entity Extraction and Linking**. Keyphrase Extraction. Automatic Topic Tagging and Classification in multiple languages.

Deep analysis of content to extract Relations, Typed Dependencies between words and Synonyms, enabling powerful context aware semantic applications.

Rapidly extract custom products, companies and build problem specific rules for tagging content with custom categories.

Generating High Ranked Entity Metadata

1. **Taxonomy based Classification**.

Algorithms for classifications according to Industry standard Taxonomies such as IAB/IPTC.

A highly comprehensive taxonomy having more than 2000 categories which can suit almost any industry vertical.

A highly comprehensive taxonomy suited for e-commerce needs.

Framework which can classify according to above pre-prepared taxonomies.

It can train models based on clients data and taxonomies.

Framework can take custom taxonomy list as input and give accurate classifications according to custom List.

1. **Article Summarizer (Summarize News Articles)**
2. **Hierarchically summarize a large cluster of similar news articles.**
3. **Sentiment Analysis Engine**

Does Sentiment Analysis of Text beyond - Positive, Negative, Neutral, considers emotions in each quadrant (anger, disgust, fear, happiness, sadness and surprise) to classify text.

Have built an Application which mines sentiments, opinions, and emotions from web and generates data sets for training models for sentiment analysis.

1. **Emotion Persona Derivation** from User Personas/Faces using Deep Learning.

Twitter corpus, Sentic Net is used.

7) **Demographic properties** derivation (Near Real time Layer + Batch Aggregation Layer)

8)**User** **Color Wheel Persona** derivation using user demographic properties + (persona vector, color affinities based on third party datasets – wikipedia, twitter, conceptnet)

9)**User Personality detection engine**.

**Big-Five personality traits**

Extroversion

Neuroticism

Agreeableness

Conscientiousness

Openness

10)**Vernacular** support.

Cleaning and Ranking Wikipedia categories corresponding to Wikipedia Article Titles.

## **User Profiling**

These Data Points are in different domains -

1)Location

2)Platform Based

3)Time Unit

4)Behavioral Segments

5)Topic Segments

**Publisher Analytic Segments**

**Engagement Time Based Segments**

1)Segment Engagement Times

2)Topic Engagement Times

3)Section Engagement Times

**Session Based Segments**

1)Session count

2)Time Since Last Session

3)Session Engagement Time

4)Loyalty Nature

5)Visitor Count.

**Machine Learning based Segments**

This covers different domains Personas

1)Emotion Persona

2)Affinities based on Creative Persona

3)Refined Demographic Persona

4)Affinities of Topic and Segments present in User Persona with respect to campaign segments.

5)Click Segments Derived After Analysis of Clickstream. This includes boolean click segment, click=true will be set in User persona, and this Key value pair will be set up in AdManager as one of the targeting parameter.

This includes Click probability of multiple CTR models for that User Persona. (This complements AdManager CTR model)

Details and Data points corresponding to above segments have already been provided and detailed in the presentation.

These data points undergo enhancement and preprocessing before being populated into CTR model (based on both low and higher order feature interactions)

**Persona Vector Generation Algorithm**

It uses deep interest network Activation Layer to learn weights of different segments present in persona vector. DIN purposes to introduce an attention method to learn from the sequence(multi-valued) features on which a local activation unit is used to get the activation score between candidate campaigns and historic segments (which a user visited).  
Conventional DIN uses impression, clicks, conversions to get an activation score for the segment.  
Advance algorithm takes care of other parameters like publisher segments - engagement time, session based parameters, ML segments to get an activation score for a segment.  
Persona vectors consists of individual segments and topic vectors.  
Weights of this individual segments, are determined by the activation score and segment vectors present in persona are added to get a segment vector.  
Similarly weights of topics are learned in above way, to get a weighted topic vector of a user.

**Analytic Engine**

* Built a Full Data Studio from scratch where a publisher corresponding to any industry can do comprehensive website analytics.
* Algorithms do comprehensive session analysis of users, gives total views, unique visitors, engagement Time.
* New, Returning, Loyal Users are determined, Devices which constitute most of the traffic.
* Custom Reports can be pulled very similar to Top Website Analytics products such as Google Analytics.
* Data can be combined in N dimensions and reports can be seen. Say for example - Report of Male Users From Gurgaon owning Samsung mobile phones and are Cricket Lovers.
* Data Studio shows Full Taxonomies which are contributing best to the website.
* Full stories can be derived by Adding Filters to data such as Cricket Lovers, Male, Samsung Mobile phone.
* Studio not only gives visualisation. It is built on top of Analytic Engine which forms the core. Powerful API framework in which you can add filters and mine data of your website according to different dimensions and derive stories.
* One can integrate these APIs right into your application which want to utilise the power of DMP.
* Analytic Engine offers topic based profiling. Cookies are mapped to topics as described above which are more granular than Interest Persona.
* It offers Full 360 User profile as well.
* One can supply cookie\_id/userid to API and it returns a Full 360 profile of industry which one can utilise in different Applications.
* It can be custom tailored to any industry vertical, publishers, ecommerce websites offering brand, seller, catalogue, Product Topic, Product analysis. Authors, Topics, Section, Article Analysis of News Websites.

## **Recommendation Engine**

• Topic/Interest Persona generated for User is used to recommend Articles to user by using K nearest Neighbours / Vectors Algorithm. Also, Article is recommended on basis of current Article view and search results can be personalised.

## **Marketing Automation Engine**

* Full Marketing Automation Engine. It is very similar to top products like Adstage.
* Through this one Set up Campaigns in Facebook, Google, Native Channels.
* One can do cross channel Look Alikes across different channels and retargeting across different channels.

**SSO Engine with cross domain support with full APIs (Currently hosted on my personal hosting account)**

API Documentation + User Persona Integration submitted.

News comments ingestion in SSO for analyzing user sentiment based persona for personality detection algorithm

**Blockchain Engine (Currently hosted on my personal hosting account)**

Ethereum layer 2 chain framework (gas less fast transactions), transfer gateways, allegro big cache, DApps, cryptocurrency/wallet design, online rewards management, trading card game, interoperable blockchains (transfer gateways), Go based smart contracts, Solidity based smart contracts, decentralized chat server application, Adserver application, social network application, smart contract based ad fraud prevention,(impression, click event trackers storage in blockchain via Go based smart contract transactions (scaled using layer 2 chains), Block Explorer, Block/Transaction Level DB database, RPC API layer for data fetch from blockchain, DPOS consensus engine, coin minting policies, Whitelisting smart contracts). Tendermint based Full implementations available for demo with full source code. Blockchain networks used – Ethereum, Tron, Binance. No third party libraries costs (only gas fees for Ethereum, tron, binance based transactions).

**User + Creative Knowledge Graph Module**

User segment and creative entities affinity derivation updated on new creatives detection on Aajtak/Indiatoday.

## **Reporting Engine (Not Hosted at present)**

* Reporting Engine mines campaign Data from different Channels via their APIs like Google Adwords API, DoubleClick Bid Manager API, Facebook Ads Insights.
* Different Reports such as Campaign Performance, Demographic Performance, Segment Performance, Location Performance and more are pulled via APIs.
* This data is aggregated in data store.
* Campaign Optimisation Engine/Targeting IQ Analysis this data and derive insights like which Targeting parameters are leading to KPIs at lowest cost such as which interest segments are leading to conversions at lowest cost, which cities are leading to conversions at lower cost.
* All these core components are plug and play and can be integrated in Applications via APIs.

## **Adserving Solutions (Not hosted at present)**

* Adserver with different Targeting options - Geo-Targeting, Device based Targeting,

Carrier based Targeting, Demographic based Targeting, Persona based Targeting, Topic Persona based Targeting. Support for Custom Input / Emotion based Impression Tracker to collect data and make 360 emotion personas.

* Frequency Capping. Prioritising Campaigns based on KPIs such as CPM, CPC, CPR, CPCon.

Implemented support for retargeting particular cookie pool.

* One can also do hybrid of Topic and Interest based Targeting.

**Integration**

**1)AdManager**

i)User key value pairs synchronizing javascript via GPT APIs.

Can also be integrated with Pubmatic Openwrap.

(Key value sync with GPT Tag latency measurements in case of mobile devices (android, iphone) (for mobile web), Desktop

ii)Third Party Tracking

Two trackers that will be ingested.

This will help to generate Data Transfer Files for two events - Impression and Clicks, generated on Google Servers locally.

Details of Click Tracker, Parameters Click Trackers will ingest.

1)Creative Url

2)Creative Size.

3)Campaign Id.

4)User Id.

5)User Persona Signals.

6)Page Signals from which clicks came.

and other general parameters derived from Request Headers

Details of Impression Tracker, Parameters Impression Trackers will ingest.

1)Creative Url

2)Creative Size.

3)Campaign Id.

4)User Id.

5)User Persona Signals.

6)Page Signals from which clicks came.

iii)Batch Upload UserIds (synchronized with google userids using cookie syncing pixel) with segments to AdManager

iv)Batch Upload GAIDS/IDFA with segments to AdManager

Ad manager limits

Characters per key 20

Characters per value 40

Active values per network 2,500,000

Active keys (dynamic and pre-defined) per network 200

Active values (dynamic and pre-defined) per key 100,000

Reportable key-values 30,000

Key values Character Limit which can be supplied to AdManager – 15000 characters.

This Character L imit is strictly adhered by javascript which supplies KVs to GPT Ad Manager by using Persona Size Optimisation Algorithm.

2)**DMP Adwords Integration** also available.