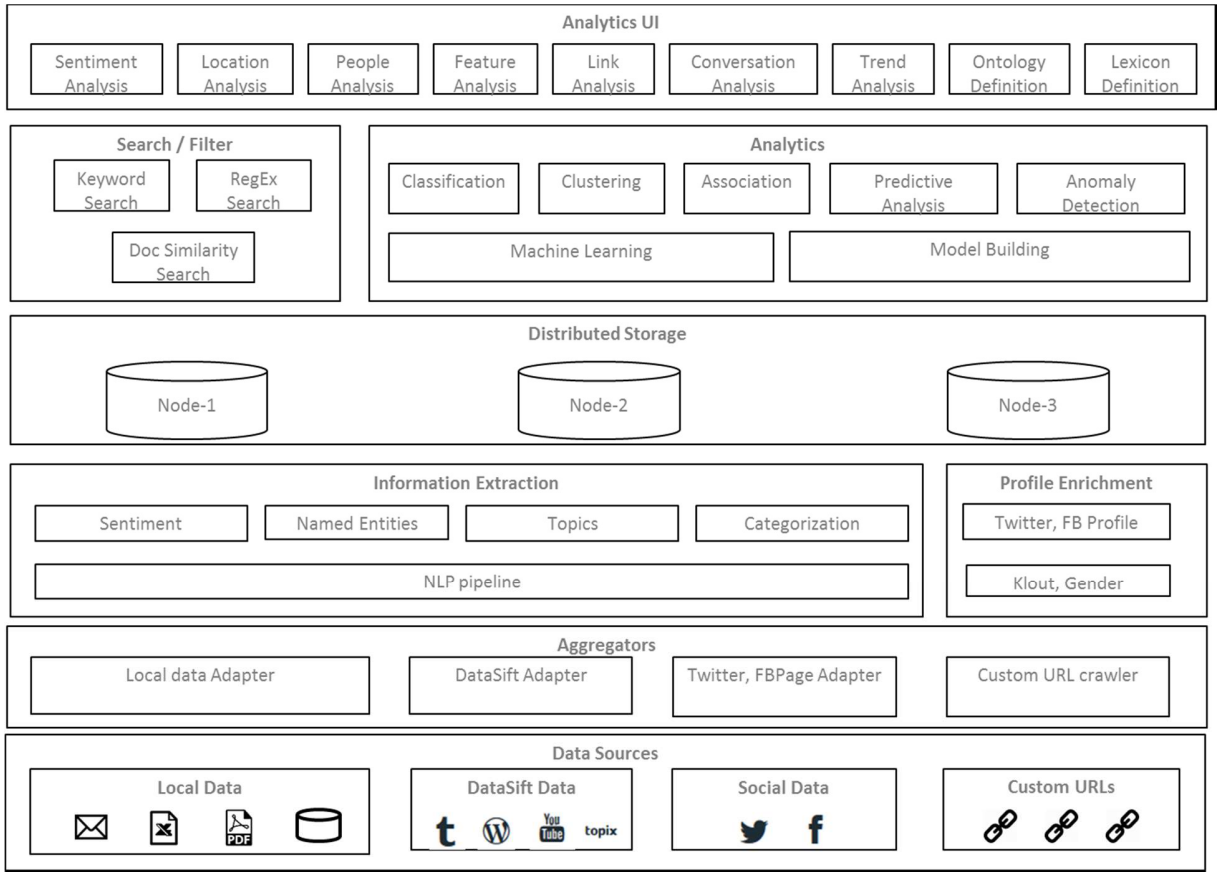
Solution Architecture

The below diagrams shows the propose solution architecture for full-filling the web mining and data analytics requirements.

Layered View



Data Sources: Various internal, external, social and web URL data sources to pull data from

Aggregators: Aggregators helps you to pull data from multiple data sources. In-built adapters to pull data from internal sources like email, excel, pdf, RDBMS and from social sources like Twitter, Facebook pages. It also includes adapter for 3rd party data aggregator like DataSift, Webhose.io etc

Further a custom URL crawler is provided which will crawl and scrape content from any web site on a scheduled basis.

Information Extraction: Components to extract Named Entities, Sentiment, Topics etc; It uses eMudhra Prism's proprietary NLP engines pipeline comprising of following modules to process content.

- Sentence tokenization
- Parts of Speech
- Chunking
- Named Entity Recognition
- Concept Extraction
- Subject Verb Object Extraction
- Dictionary Based Extraction (of specific tags)
- Sentiment Analysis

Profile Enrichment : Custom components to enrich a social users profile like with gender, location, education, college, interests, likes, Klout score etc;

Distributed Storage: Data pulled from various data sources in heterogeneous form is stored in a big data based distributed storage platform and will be consumed further for analytics and querying

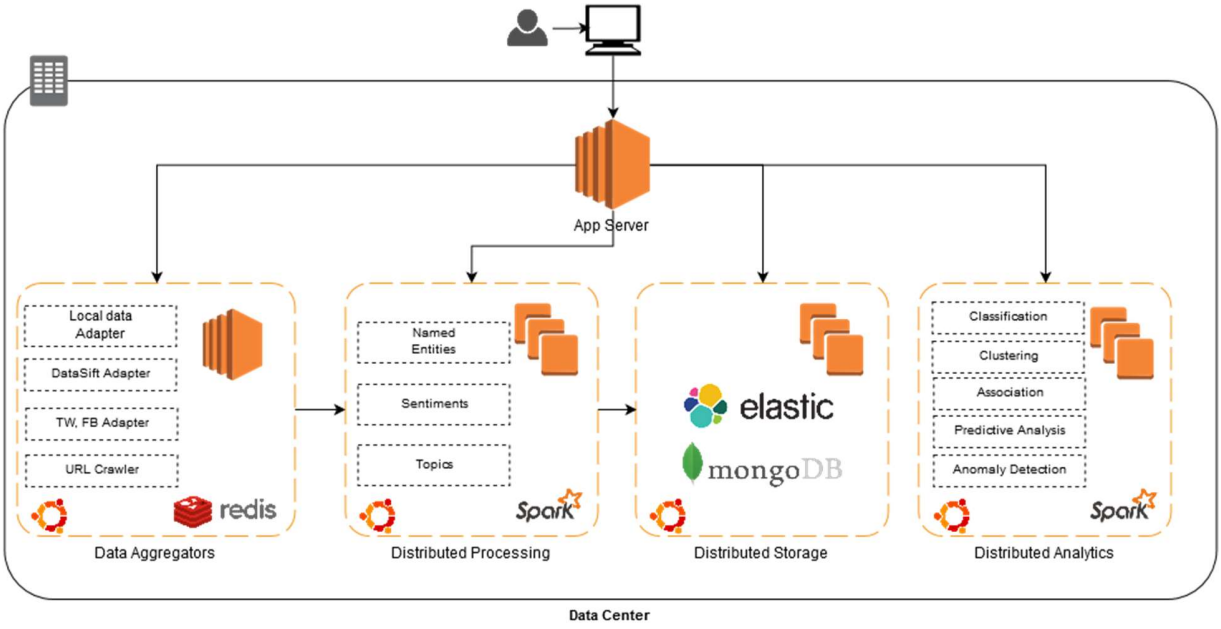
Search / Filter: Retrieve data by searching both structured and un-structured content and apply various column based filters.

Analytics: Perform user defined classification, clustering, predictive analysis etc; by training data with various machine learning algorithms and creating and running models.

Analytics UI: Provides interface with various pre-defined analytics and also provides ability to do user defined analysis by creating & running custom machine learning models.

D3 library will be used for chart visualization.

Deployment Architecture



Data Aggregators: All the data adapters will be running on this box. An in-memory high performance store like Redis is used to temporarily store high volumes of data received from each adapter.

Tools such as Scrapy, Selenium are used for custom URL crawling.

Distributed Processing: All the data is processed for extracting Named Entities, Sentiments, Concepts etc; using a distributed computing platform like Apache Spark to handle large volumes & high velocity of data.

Distributed Storage: Raw data from aggregators and processed data from the NLP engines is further stored into a permanent data store. Big data and distributed stores like MongoDB and Elasticsearch are used for this purpose. It also helps in during any custom analytics using Spark.

Distributed Analytics: Runs user defined machine learning models on a distributed platform.