DATA ANALYSIS AND VISUALISATIONS OF

INDIAN PREMIER LEAGUE

Nithin Kankanti  
 Computer Science  
 University of Colorado Boulder  
 Boulder Colorado United States  
 nika9944@colorado.edu

Mohan Dwarampudi  
Computer Science  
 University of Colorado Boulder  
 Boulder Colorado United States  
 modw1211@colorado.edu

Sankaranarayanan  
 Computer Science  
 University of Colorado Boulder  
 Boulder Colorado United States  
 sana6469@colorado.edu

**INTRODUCTION**

The Indian Premier League is a professional Twenty20 cricket league in India contested during March or April and May of every year by eight teams representing eight different cities in India. In this project, we want to present some cool, interesting fun facts, records and results of the Indian Premier League..

Given the amount of data collected over the years, this is a perfect scenario to do data analysis, mining. Due to the abundance of data, it enables us to explore the various aspects of the data pipeline as well and how it can scale. This can be used in building a distributed data analysis/mining system..

A huge amount of data indicates that Data Visualizations can also be utilized to generate useful insights. This can also be used as an example of knowledge mining. It can also be used as a feedback mechanism for the players which will help them in improving the performance.

**KEYWORDS**

Data Visualization, Distributed Data Analysis, Data Mining

**RELATED WORK**

Previous work in this domain involved doing Fuzzy Clustering on IPL data for data mining.Fuzzy clustering is an extension of the cluster analysis, which represents the affiliation of data points to clusters by memberships. Introducing fuzziness to clustering gives us the flexible representations of substructures of the data set.There are different shapes of cluster centers and prototypes. Most of them conduct clustering in accordance with similarity or dissimilarity derived from distances from the centroid of the cluster of the data points.

Other work in this domain involved doing Predictive Analysis on the dataset. Examples include analysis on the dependency between winning the toss and batting and whether the decision was successful in the end. It also includes trivial analysis on how strategies can be implemented given the pitch and weather conditions. Some examples include analysis of players in the teams and the correlation between the type of pitch andstrengths of the players in the team.

**PROPOSED WORK**

In this project we are going to build an end to end ETL pipeline to mine the data to extract valuable insights for better visualizations. ETL Processing will be done via amazon web services data pipeline due to amount of input data.

The ETL Pipeline that we came up with is as follows:

**A) EXTRACT:**

Data has been sourced from multiple areas

* Scrapping popular cricketing websites
* Scrapping wiki
* Google API – Geo points
* Kaggle competitions.

All the data is further loaded into s3. Raw data received in S3 is pushed into AWS DynamoDB.S3 event invokes AWS Lambda which does the data parsing before it is rested in DynamoDB. Three tables – Player, Matches, Deliveries. Data in DynamoDB acts a source of truth for all the further operations.

**B) TRANSFORM:**

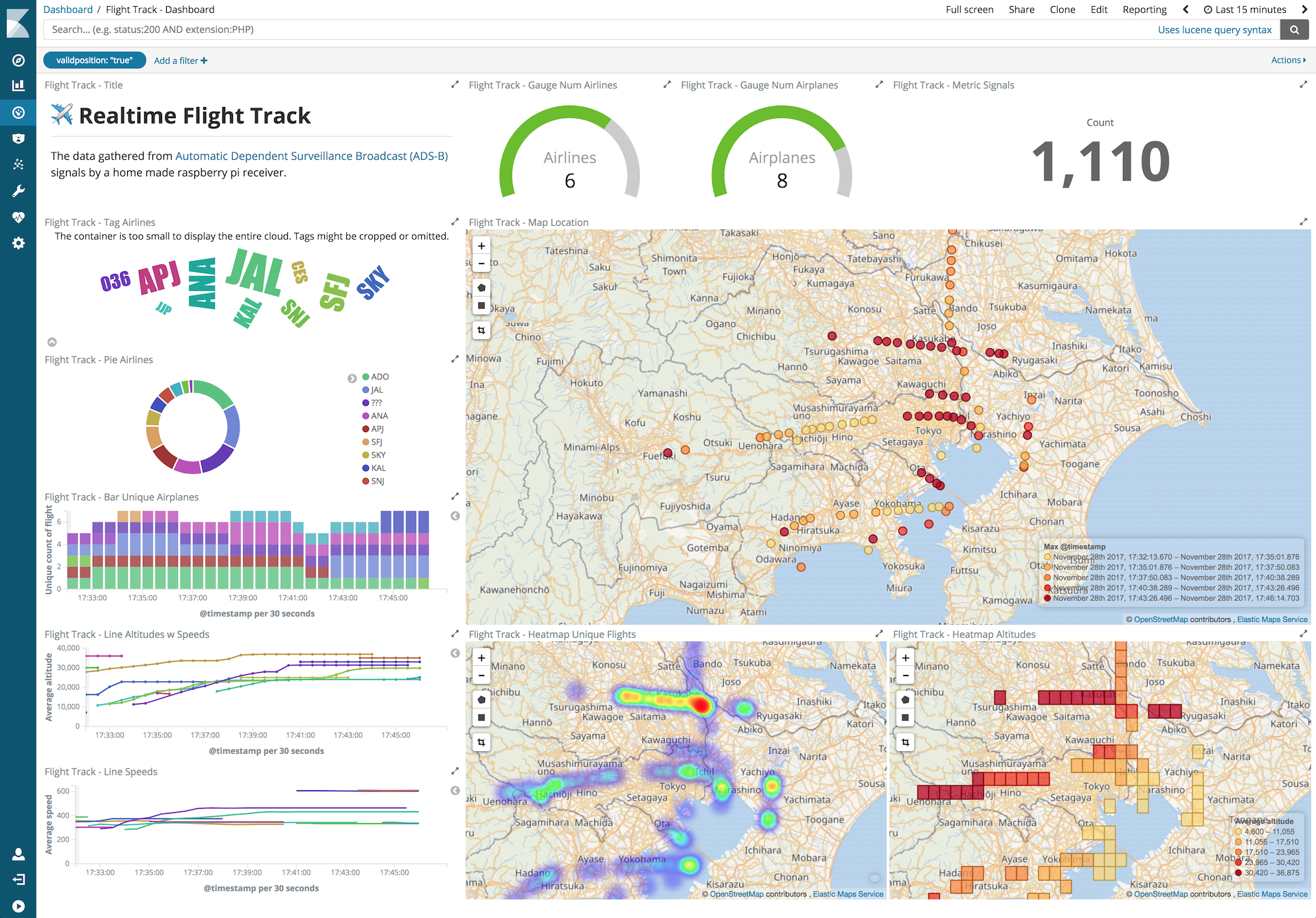
All the semi parsed data is transformed into a meaningful entity - JSON. Triggers on DynamoDB would invoke AWS Lambda whenever a new entry is added into DynamoDB. AWS Lambda transforms the data into a meaningful pattern, which are further loaded into elastic search cluster. AWS Lambda would fetch additional geo data through Google API. AWS Lambda uses Redis for a quick key-value mapping lookup.

**C) LOAD:**

Elasticsearch indexes all the incoming data from lambdas. Data on elastic search is split on the nodes in the cluster. All the three formats of the data are stored in different indices namely matches, players and deliveries.

**EVALUATION:**

* We will be using kibana to generate dashboards separately for each of the index.
* We can apply filters on a visualisation to come up with the required metrics.
* We will be generating different visualizations such as Geo Tagging, Heat Maps, Custom Metrics, Pie charts.
* Our project end product will be looking similar to the one mentioned below.



**MILESTONES:**

* Data Extraction from different sources and storing into S3.
* Creating AWS lambda functions to parse the data before loading into DynamoDB.
* Transforming the data into JSON Format.
* Fetching geo data through Google API.
* Transformed data is loaded into elastic search cluster.
* Indexing the incoming data from lambdas with elastic search.
* Loading the data into three different tables namely Matches, players and Deliverables.
* Generate visualisations separately for each of the tables with the created indices using Kibana.
* Create the dashboards with the already generated visualisations.
* Use NGINX to host the created dashboards.

**ACKNOWLEDGMENTS**

Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

**REFERENCES**

[1] Patricia S. Abril and Robert Plant, 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan, 2007), 36-44. DOI: <https://doi.org/>10.1145/1188913.1188915.

[2] Sten Andler. 1979. Predicate path expressions. In *Proceedings of the 6th. ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages (POPL '79)*. ACM Press, New York, NY, 226-236. DOI:https://doi.org/10.1145/567752.567774

[3] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. DOI:https://doi.org/10.1007/3-540-09237-4.

[4] David Kosiur. 2001. *Understanding Policy-Based Networking* (2nd. ed.). Wiley, New York, NY..