**Advertisement Data analysis**

Domain : Advertisement

No. of workers : 9-10 (overall on project 30+)

Client : Advertisement Domain client

Problem to solve :

Client wanted to build an end to end pipeline

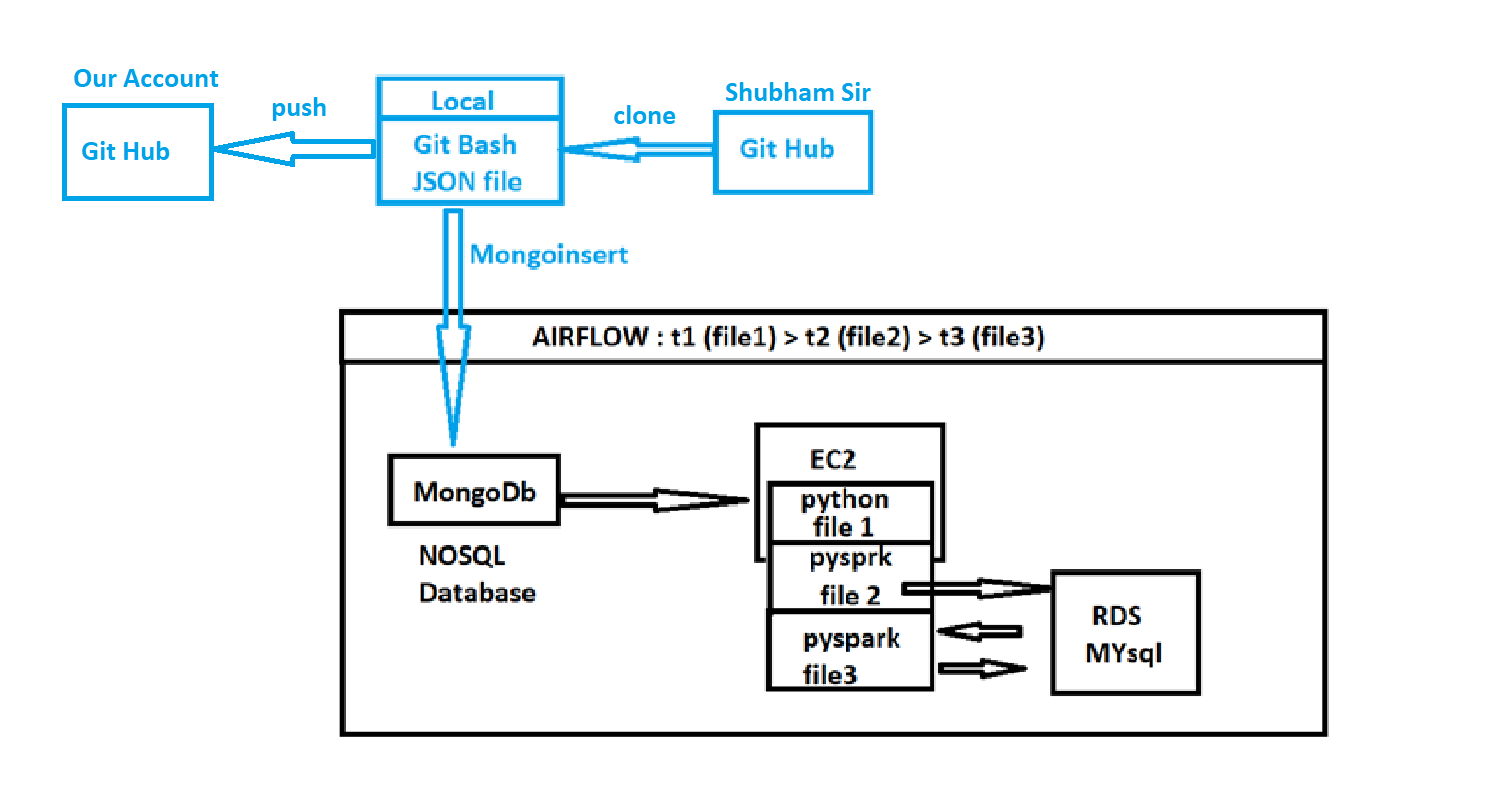
* Developing the ad campaigns
* Reviewing user clicks
* Developing

Solution :

The client was in domain, it had daily generating data from 1000s of websites and all the websites are saving are saving the data in mongo atlas (a cloud database). The job was get the data from mongo db, load into my sql. Transform it according to business logic and create the tables in mysql, from there the data analytical team used to pick the data.

1. The data was JSON and it continuously appended in a mongo db table
2. From their an airflow job was schedculed daily which used to take the data from location, append one days data in mysql
3. From there spark job was schedules

**Project Pipeline :**



Mongodb : 1 day data

Mysql : Store daily data upto 2 years

Spark : Read All data from Mysql then process it and it will store on Mysql in different table

EC2 : Airflow + Python 3.8 + Spark 3.1.2