**DOKUZ EYLUL UNIVERSITY**

**ENGINEERING FACULTY**

**COMPUTER ENGINEERING DEPARTMENT**



**Social Media Analyse Using Hadoop**

**Report#2**

**SociAnalyzer**

2008510057 Ferhat YENIGUN([ferhat.yenigun1@gmail.com](mailto:ferhat_yenigun@hotmail.com))

2009510107 Halil SATIK([halilsatik@outlook.com](mailto:halilsatik@outlook.com))

2010510004 Bilal AKSOY([bilal\_roda@hotmail.com](mailto:bilal_roda@hotmail.com))

2010510046 M. Hanifi KOC([muha\_koc@gmail.com](mailto:muha_koc@gmail.com))

### 

### Advisor: Asst. Prof. Dr. Adil ALPKOCAK

**Function 1**

Data Format Variation

**Description:**

There are a lot of social media sites, like Facebook,Twitter, Instagram etc. Some of them will be used in this project. Each of these platforms use different data format. This case force to make a separate process for each data format.

We will try to develop a parser method (like XML parser, JSON parser) for each them.

**Input:**

XML, JSON, line-based etc. unstructured or semi-structured data

**Output:**

Structured data that is convenient for storing database

**Function 2**

Gathering Data

**Description:**

Getting data of Social web sites like twitter, facebook are not free, getting data from them cost a lot and %1 part of data is free but this is not sufficient for analyzing and it will not feasible due to this in January, we are going to make a deal with twitter for taking data.

**Function 3**

Data Store

**Description:**

Keeping large data in relational database management systems generally causes troubles, keeping large data in ordinary database brings large expense. Consequently, to deal with this problem nosql based system should be used.

**Input:**

Structured data

**Output:**

Data stored on the database

**Function 4**

Query Optimization

**Description:**

Creating suitable queries for the data that users input are costly. Because user’s input what they want and they are not aware of how things happen at background. For dealing with this problem, words that users input should be in an optimized form.

**Input:**

User’s raw query

**Output:**

Simplified and optimized query

**Function 5**

Representation of Results

Results would be different according to queries of users. Our purpose is changing these results into understandable. About this subject, Necessary tools will be used.

**Input:**

User’s result

**Output:**

Result’s graphically representation