BİL401

Project Proposal Report

Kerem Elma - 211101079

Mehmet Eski – 221104095

1. **Project Purpose**

Analyzing performance differences of big data with processing processes by comparing XML, JSON, Avro and Parquet formats. It is an important issue to compare the processing of different data formats and see which one is more convenient to use. With using Apache Spark, analyzing read and write speeds and processing times can be analyzed.

1. **Source of Dataset**

The dataset titled "Air Quality of New York" is available on the website "[https://data.gov](https://data.gov/)." The dataset can be accessed from the link below;

* <https://data.cityofnewyork.us/api/views/c3uy-2p5r/rows.json?accessType=DOWNLOAD>

1. **Description of Dataset**
   1. **Domain**

Dataset contains information on New York City air quality surveillance data. The attributes related to air quality measurements.

* 1. **Columns**
  + **Unique ID:** Unique record identifier
  + **Indicator ID:** Identifier of the type of measured value across time and space
  + **Name:** Name of the indicator
  + **Measure:** How the indicator is measured
  + **Measure Info:** Information (such as units) about the measure
  + **Geo Type Name:** Geography type; UHF' stands for United **Hospital Fund neighborhoods;** For instance, Citywide, Borough, and Community Districts are different geography types
  + **Geo Join ID**: Identifier of the neighborhood geographic area, used for joining to mapping geography files to make thematic maps
  + **Geo Place Name:** Neighborhood name
  + **Time Period:** Description of the time that the data applies to; Could be a year, range of years, or season for example
  + **Start Date:** Date value for the start of the time\_period; Always a date value; could be useful for plotting a time series
  + **Data Value**: The actual data value for this indicator, measure, place, and time
  + **Message**: Notes that apply to the data value; For example, if an estimate is based on small numbers we will detail here
  1. **Size of Dataset**

**•** The dataset contains 19,002 rows and multiple attributes

1. **Platforms and Tools**

**Processing Big Data:** Apache Spark

**Data Formats:** XML, JSON, Avro, Parquet

**Storage:** HDFS

**Veritabanı:** Apache Hive

**Frameworks:** PySpark

**Cloud Systems:** Google Cloud(can be changed)

**Hardware:** Local machine