Assignment 2: Big Data Pipeline with Spark

Objectives:

- Design streaming and/ or batch processing systems to collect and analyze real time data
- Use Apache Spark and other relevant software to implement a data pipeline.

Note*: Lectures and labs will cover the required concepts and implementation technologies.*

Data pipelines constitute the refineries of data: they efficiently collect, process, and transform raw data into actual value. Goal of this assignment for you is to implement a data pipeline to process a dataset of your interest and extract value or insights from it.

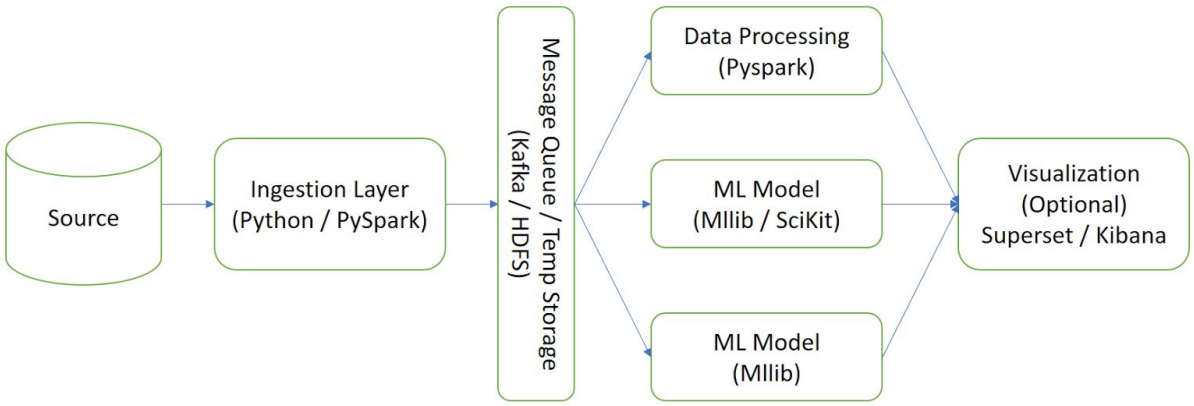
Assignment:

For your selected dataset write **three queries** which are interesting! Additionally, build and deploy a machine learning model on top of your data pipeline. Finally, you can tie it together with a nice dashboard with a visualization tool. This is an optional step, but is considered for bonus points. Please note as this is a data pipeline, naturally all your submissions should be executables and not running in an interactive environment like notebook.

The following technologies are recommended but not mandatory:

Apache Spark, Hadoop Distributed File System, Kafka, Hive, Superset, Elasticsearch, Kibana

With the technologies above you need to implement the following architecture:



Submission Guideline

Deliverables (See Deliverable guidelines)	A report in PDF or Doc and source codes A demonstration A presentation
Group Size	3
Deadline for Report and Source code Submission	2017 December 4
Presentation Time	2017 December 5
Demonstration	Presentation and demonstration may be combined.

Deliverable Guidelines

Report	Content : ● Overview of the report ● Design and implementation of data pipeline ● Queries for generating Insights ● Application of machine learning ● Information about each student in the group (full name and student numbers) ● Contribution of each student (Appendix) in few sentences Size : less than 10 pages
Demonstration / Presentation	● Demonstrate execution of data pipeline ● Share your insights! ● Show off the visualization dashboard (if any) Time : Less than 15 minutes
Source code	All source files (Do not upload binaries/compiled files, and datasets)

Marking Scheme

Implementation & Architecture	20%
Insights / Queries	20%
Machine Learning Model	15%
Wow Factor	15%
Demonstration & Presentation	30%

Marking Criteria per Category

Category	What we are looking for
Implementation & Architecture	Justified choices of technologies to enable a more robust and scalable data pipeline.
Insights / Queries	Queries with medium complexity using built-in functions or user-defined functions.
Machine Learning Model	Simple implementation of machine learning model.
Wow Factor	Show something cool! Perhaps visualization or a nice use case of machine learning / advanced analytics.
Demonstration & Presentation	Well-captured overview of the solution backed up by a nice demo.