Project Assignment

Big Data Project

Delivery at 25/05/2025 23:59

Project Overview

Big Data Analytics
(MAA/DSAA)

Students (in teams of 3 to 5 by 28 of February) are expected to select a Big Data problem, process and analyze large datasets using appropriate Big Data tools, and present their findings.

Key Requirements

1. Problem Definition: Identify a relevant Big Data problem. Be creative.
2. Data Collection & Preprocessing: Obtain, clean, and pre-process large datasets.
3. Big Data Processing: Use Apache Spark Modules (SQL, MLlib, or Streaming).
4. Data Analysis & Visualization: Apply machine learning techniques, statistics, or BI tools.
5. Results & Insights: Present findings through dashboards, visualizations, or reports.
6. Project Presentation: Deliver a 7–10 min talk, followed by Q&A.

Potential Project Ideas [Select only one or another topic you like]

Pick a single bullet point to work on in your project in a single domain area.

1. Business & Finance

- Stock Market Prediction using Big Data & Spark MLlib
- Fraud Detection using real-time transaction streams (Kafka & Spark Streaming)
- Customer Segmentation using Clustering in Spark MLlib

2. Healthcare & Environment

- Disease Prediction & Analysis from healthcare datasets
- Air Pollution & Climate Change Trends using Big Data visualization
- Biodiversity Monitoring from satellite or sensor data

3. Social Media & E-commerce

- Sentiment Analysis of Twitter/X Data using NLP in Spark
- Recommendation System for e-commerce using Graph Analytics
- Influence Analysis in social networks using GraphFrames

Note 1: Be aware that a “boring” dataset makes for a poor project. This project is supposed to give you the capabilities to work with Spark in its full breadth. Then, make sure that your choice of dataset allows for suitably advanced explorations.

Note 2: Be also aware that some of the project listed here or chosen by you might require the installation of external libraries that may not have been covered explicitly in class. This is not a reason to abdicate from the responsibility of pursuing your project analytically to the end.

Deliverables

Technical Report (3 5 pages strict limit!)

- Problem statement (Why?)
- Data processing workflow (How?)
- Algorithms used (How?)
- Results & insights (What?)
- Challenges & future improvements (What’s more?)

Source Code (Databricks Notebook)

- Notebook – Python

Source code that adheres to the following criteria will score better in the evaluation of the code grading component:

- Well-commented and pertinent code blocks and markdown cells for explanations

- Demonstrable abstraction from code blocks developed in class, i.e., very (!) limited re-use of code

- Adherence to Big Data principles, i.e., the avoidance of pure Python code for processing datasets and well-dosed/justified usage of Python libraries for visualization (make sure to use Databricks' built-in chart capabilities or external tools suitable for larger datasets)

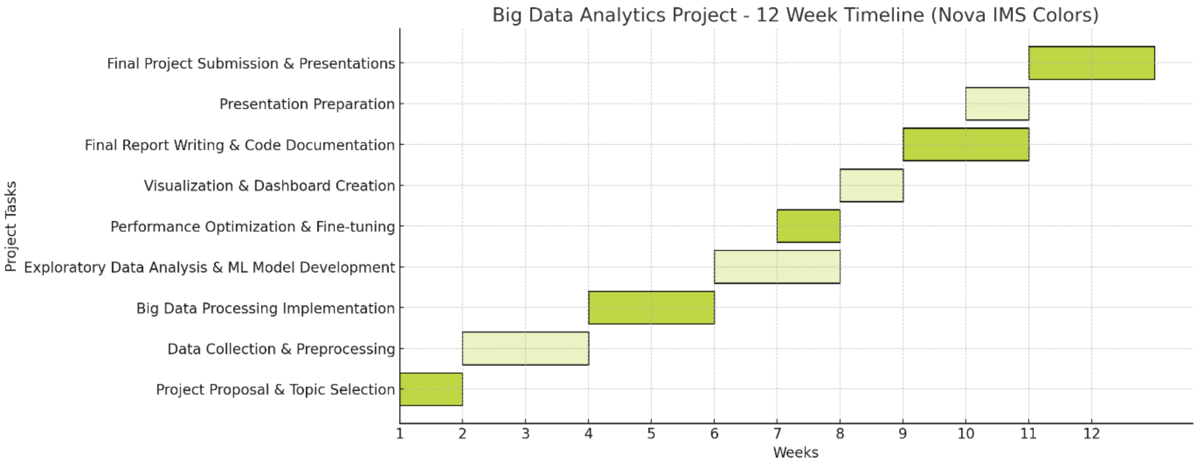
- Demonstration of the spectrum of the Spark API: Usage of RDDs, DataFrames and Inline-SQL

Presentation (7 10 min, slides)

- Explanation of the problem

- Key findings & impact

Make sure to include a project time in a Gantt chart (see an illustrative example below).



Grading Criteria

Category	Points
Problem Definition	10 pts
Data Preprocessing	20 pts
Big Data Processing	25 pts
Analysis & Insights	20 pts
Visualization & Presentation	15 pts
Q&A & Peer Engagement	10 pts
Bonus for Streaming [Optional]	10 pts
Bonus for GraphFrames [Optional]	20 pts
Award for Outstanding Writing	10 pts
Total	100 pts