Project Assignment

Big Data Project

Delivery at 25/05/2025 23:59

Project Overview

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 preprocess large datasets.
3. Big Data Processing: Use Apache Spark Modules (SQL, MLlib, or Streaming).
4. Data Analysis & Visualization: Apply machine learning, 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]

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 GraphX

Deliverables

Technical Report (3 5 pages strict limit!)

- Problem statement
- Data processing workflow
- Algorithms used
- Results & insights
- Challenges & future improvements

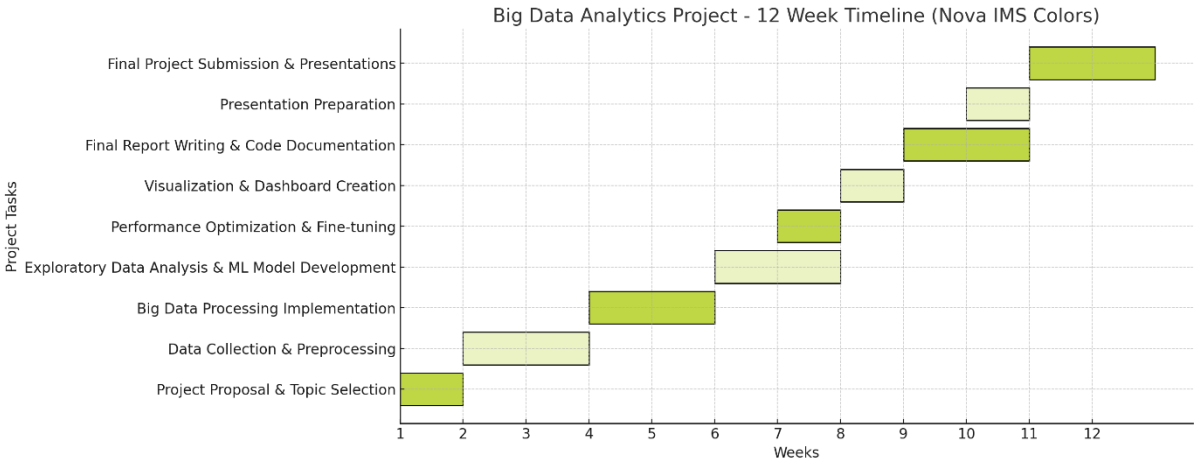
Source Code (Databricks Notebook)

- Notebook - Python

Presentation (7 10 min, slides)

- Explanation of the problem
- Key findings & impact

Do not forget 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 GraphX [Optional]	20 pts
Award for Outstanding Writing	10 pts
Total	100 pts