

Phase 1: Ideation & Dataset Discovery (Week 1)

Goal: Define what problem you want to solve and choose the right data.

- Brainstorm real-world problems in the chosen domain (healthcare, finance, retail, etc.)
- Shortlist 2–3 relevant, large, open-source datasets
- Evaluate datasets for completeness, size, structure, update frequency, and license
- Finalize topic in discussion with faculty and align on project scope

Domain Exploration Approach

- Select a domain you like: healthcare, education, sports, environment, e-commerce
- Narrow down to a specific challenge (e.g., patient wait times, student dropout prediction)

Reverse Engineering

- Look at existing dashboards or AI tools and try to recreate or enhance them
- Study Kaggle competitions and simplify for your own scope

Personal Frustration Model

- Think of something that annoys you in daily life (e.g., food delivery delays, expensive cab fares)
- Ask: "Can this be solved using data?"

Key Tools:

- Kaggle Datasets
- AWS Open Data Registry
- UCI Machine Learning Repository
- Google Dataset Search
- data.gov / data.gov.in
- World Bank Data
- OpenWeather API

Before finalizing a dataset, ask:

- Is the dataset large enough for real Big Data tools (min 5-10 GB recommended)?
- Is it structured, semi-structured, or unstructured (CSV, JSON, XML)?
- Is the data recent, relevant, and complete?
- Are there enough dimensions to extract insights or train models?
- Are there data quality issues? Missing values? Outliers?

Phase 2: Problem Framing & Objective Setting (Week 2)

Goal: Create a focused and impactful problem statement.

- Frame a clear problem statement: “What problem are we solving?”
- Define objectives (e.g., Predict X, Automate Y, Visualize Z)
- Identify target audience or business use case
- Break it down into measurable milestones

Deliverables:

- One-line Problem Statement
 - Project Goals & Objectives
 - High-level Use Case Diagram
-

Phase 3: Architecture Planning (Week 3)

Goal: Visualize how all technologies will connect to build the solution.

- Decide input/output flow and storage design (e.g., raw data in S3, processed data in Glue)
- Design cloud architecture using services: EC2, S3, RDS, Lambda, Glue, SageMaker, etc.
- Draft architecture diagram including: data flow, tools used, automation points

Deliverables:

- Tech Stack Overview
 - Architecture Diagram
 - Tool Justification Sheet
-

Phase 4: Data Engineering & Preprocessing (Week 3–6)

Goal: Make raw data usable and insights-ready.

- Clean, filter, and transform large datasets using Pandas or PySpark
- Create partitions, handle schema drift, manage incremental updates
- Store final cleaned dataset in AWS S3, RDS, or Hive

Deliverables:

- Cleaned and Processed Dataset
 - Data Dictionary (with column-wise explanation)
 - ETL Job Scripts or Workflows
-

Phase 5: Environment Setup & Workflow Automation (Week 6)

Goal: Set up the infrastructure and repository for collaboration.

- Set up GitHub Repository with branching model
- Configure CI/CD with GitHub Actions (data ingestion trigger, notebook validation, etc.)
- Use CloudFormation or Terraform to deploy AWS infrastructure (S3, IAM, Glue, etc.)

Deliverables:

- GitHub Repo with Branches and Workflow
 - Infrastructure as Code (IaC) templates
 - Connection and Access Setup (IAM roles, secrets)
-

Phase 6: Analysis / ML Model Building / AI Integration (Week 7–8)

Goal: Add intelligence to the data.

- For analytics projects: perform EDA, identify trends, generate insights
- For ML: preprocess features, train/test models, evaluate metrics
- For AI: integrate AWS Comprehend, Polly, Rekognition, Lex for use case
- Implement feedback loops or automation (e.g., Lambda for model inference)

Deliverables:

- Jupyter/PySpark notebooks
 - Model metrics & charts
 - API endpoints or AI outputs
 - KPI Definitions
-

Phase 7: Visualization & Reporting (Week 9–10)

Goal: Present your findings in an intuitive and appealing manner.

- Create dashboards using Tableau, Power BI. Include filters, summaries, and comparisons that highlight business impact
- Build executive-level summary sheets

Deliverables:

- Interactive Dashboards
 - Insight Reports Storytelling
-

Phase 8: Rehearsal & Final Touches (Week 11)

Goal: Refine the presentation and prepare for evaluation.

- Clean up the GitHub repository (README, folder structure, documentation)
- Rehearse the 10–15 min presentation within your team
- Address likely questions: "What was your biggest challenge?", "How did you solve X?"
- Finalize PPT with architecture, KPIs, team roles, outcomes

Deliverables:

- Rehearsal Pitch Deck
 - Final Project Documentation
-

Phase 9: Final Presentation & Industry Review (Week 12)

Goal: Confidently present your work in front of faculty and industry expert.

- Each group presents their work to all stakeholders including the Vice Principal
- Final round of Q&A
- Online evaluation by Mr. Pradeep Tripathi (Industry Mentor)
- Feedback and improvement suggestions

Deliverables:

- Final PowerPoint Deck
 - Live Demo or Walkthrough
 - Evaluation Form Submission
-