Data Analytics and Reporting (DAR) – Combined Modular Assignment 7

Module 7 (2,3 & 4): Data-Driven Analytics and Reporting Pipeline

Duration: 4 hours

Problem Statement

Design and implement a complete data-driven analytics pipeline that seamlessly integrates multiple data sources, applies advanced exploratory data analysis and machine learning techniques, and delivers comprehensive reporting solutions. Create an end-to-end system that transforms raw data into actionable business insights through sophisticated analytics and reporting.

This assignment combines the core technical skills from handling data sources (Module 2), (EDA), Models and Techniques (Module 3), and reporting fundamentals (Module 4) into a unified solution.

Business Scenario

Choose ONE integrated business challenge:

Option 1: Customer Intelligence Platform

Context: Multi-channel retail company needs integrated customer analytics - **Data Sources**: - Structured: Customer database (SQL), transaction history, product catalog - Semi-structured: Web analytics (JSON), mobile app logs (CSV), email campaigns (XML) - Unstructured: Customer reviews, social media mentions, support tickets - **Analytics Goals**: Customer segmentation, churn prediction, lifetime value modeling - **Reporting Needs**: Executive dashboards, operational reports, ad-hoc analysis capabilities

Option 2: Supply Chain Optimization System

Context: Manufacturing company optimizing global supply chain operations - **Data Sources**: - Structured: ERP system (SQL), inventory databases, supplier information - Semi-structured: IoT sensor data (JSON), shipping manifests (CSV), regulatory documents (XML)
- Unstructured: Supplier communications, market intelligence reports, news feeds - **Analytics Goals**: Demand forecasting, supplier risk assignment, inventory optimization - **Reporting Needs**: Performance dashboards, exception reports, predictive analytics views

Option 3: Financial Risk Management Platform

Context: Regional bank implementing comprehensive risk analytics - **Data Sources**: - Structured: Core banking system (SQL), credit bureau data, regulatory filings - Semi-structured: Market data feeds (JSON), transaction logs (CSV), compliance reports (XML) - Unstructured: News sentiment, regulatory announcements, customer communications -

Analytics Goals: Credit risk modeling, fraud detection, portfolio optimization - **Reporting Needs**: Risk dashboards, compliance reports, predictive risk scorecards

---

Assignment Requirements

Part A: Multi-Source Data Integration and Architecture (60 minutes)

Module 2: Handling Data Sources

Tasks: 1. **Comprehensive Data Source Integration** (35 minutes) - Design data architecture for all three data types: - **Structured Data**: Implement normalized relational database design - **Semi-structured Data**: Create JSON/CSV/XML processing pipelines - **Unstructured Data**: Build text processing and extraction systems - Implement ETL/ELT processes with error handling - Create data quality monitoring and validation frameworks - Design data lineage tracking and metadata management

2. **Database Implementation and Optimization** (25 minutes)
    - Set up relational database with proper normalization (3NF)
    - Implement NoSQL database for flexible schema requirements
    - Write complex SQL queries with joins, subqueries, and window functions
    - Create database indexes and optimize query performance
    - Implement data partitioning and archiving strategies

Expected Deliverables: - Complete data architecture diagram with data flow - Functional ETL pipelines for all data source types - Database schema with sample data and optimized queries - Data quality assignment and monitoring framework

---

Part B: Advanced Analytics and Machine Learning Pipeline (90 minutes)

Module 3: EDA, Models and Techniques

Tasks: 1. **Comprehensive Exploratory Data Analysis** (35 minutes) - Perform advanced trend detection and seasonality analysis - Implement multiple outlier detection methods with treatment strategies - Conduct sophisticated missing data analysis and imputation - Apply association rule mining and pattern discovery - Create correlation analysis with spurious relationship detection - Implement winsorization and percentile analysis techniques

2. **Machine Learning Model Development** (40 minutes)
    - Build comprehensive modeling pipeline:
        - **Regression Models**: Linear, non-linear, logistic with regularization
        - **Classification**: Decision trees, SVM with hyperparameter optimization
        - **Clustering**: K-means with optimal cluster determination
        - **Time Series**: Forecasting models with trend decomposition
    - Implement cross-validation and model comparison frameworks
    - Apply variable transformation and feature engineering
    - Create model performance monitoring and validation systems
3. **Advanced Model Validation and Selection** (15 minutes)

- Calculate comprehensive error metrics (MSE, RMSE, MAPE, more.)
- Implement overfitting/underfitting detection and prevention
- Create model interpretability and explanation frameworks
- Design A/B testing for model deployment validation

Expected Deliverables: - Comprehensive EDA report with statistical insights - Multiple ML models with performance comparison matrix - Model validation and selection documentation - Feature engineering and transformation pipeline

---

Part C: Integrated Reporting and Dashboard System (80 minutes)

Module 4: Reporting Fundamentals

Tasks: 1. **Report Architecture and Design** (25 minutes) - Design comprehensive report taxonomy: - **Operational Reports**: Daily metrics and KPI monitoring - **Analytical Reports**: Trend analysis and deep-dive investigations - **Executive Reports**: Strategic summaries with key insights - **Compliance Reports**: Regulatory and audit requirements - Create report templates with consistent anatomy and structure - Implement responsive design for multiple device types

2. **Top-Down Drill-Down Implementation** (30 minutes)
    - Build hierarchical reporting system:
        - **Level 1**: Executive dashboard with high-level KPIs
        - **Level 2**: Departmental performance summaries
        - **Level 3**: Detailed operational metrics and trends
        - **Level 4**: Transaction-level data with full context
    - Implement interactive navigation with breadcrumbs
    - Create cross-dimensional filtering and parameter controls
    - Add bookmarking and report sharing capabilities
3. **Bottom-Up Ad-Hoc Analytics** (25 minutes)
    - Design self-service analytics interface:
        - Drag-and-drop query builder with validation
        - Dynamic field selection and aggregation options
        - Real-time preview and result visualization
        - Query optimization and performance monitoring
    - Implement saved queries and report templates
    - Create guided analytics workflows for business users
    - Add predictive analytics integration for forecasting

Expected Deliverables: - Complete reporting system with all report types - Multi-level drill-down navigation with interactivity - Self-service ad-hoc analytics platform - Report design standards and best practices documentation

---

Part D: Integration and Business Intelligence Platform (60 minutes)

Cross-Module Integration

Tasks: 1. **End-to-End Pipeline Integration** (35 minutes) - Connect data integration with analytics pipeline - Implement automated workflows from data ingestion to reporting - Create real-time data processing and alert systems - Build comprehensive error handling and logging - Design system monitoring and performance optimization

2. **Business Intelligence Dashboard** (25 minutes)
    - Create unified BI platform combining all components:
        - Real-time data quality monitoring
        - ML model performance tracking
        - Report usage analytics and optimization
        - User activity monitoring and role management
    - Implement automated report generation and distribution
    - Create business impact measurement and ROI tracking
    - Design system scalability and growth planning

Expected Deliverables: - Fully integrated BI platform with automated workflows - System monitoring and performance optimization framework - Business impact analysis and ROI calculation - Scalability and maintenance documentation

---

Part E: Advanced Features and Innovation (30 minutes)

Enhanced Capabilities and Future-Proofing

Tasks: 1. **Advanced Analytics Features** (20 minutes) - Implement real-time analytics capabilities - Add natural language query processing - Create anomaly detection and automated alerting - Build recommendation engines for insights discovery - Integrate external data sources via APIs

2. **Platform Optimization and Governance** (10 minutes)
    - Implement data governance and security frameworks
    - Create user training and adoption strategies
    - Design backup, recovery, and disaster planning
    - Document best practices and lessons learned

Expected Deliverables: - Advanced analytics features with documentation - Data governance and security implementation - User adoption and training materials - System optimization and maintenance guide

---

Evaluation Rubrics

1. Data Integration Excellence (25 points)

- **Excellent (23-25)**: Sophisticated multi-source integration with optimal performance
- **Good (18-22)**: Good integration with minor performance or quality issues
- **Satisfactory (13-17)**: Basic integration with adequate functionality
- **Needs Improvement (0-12)**: Poor integration or significant data quality issues

2. Advanced Analytics Implementation (30 points)

- **Excellent (27-30)**: Comprehensive analytics with optimized models and validation

- **Good (21-26)**: Good analytics with adequate model performance
- **Satisfactory (15-20)**: Basic analytics with simple models
- **Needs Improvement (0-14)**: Poor analytics or model implementation failures

3. Reporting System Quality (25 points)

- **Excellent (23-25)**: Sophisticated reporting with excellent UX and functionality
- **Good (18-22)**: Good reporting system with minor limitations
- **Satisfactory (13-17)**: Basic reporting with adequate features
- **Needs Improvement (0-12)**: Poor reporting or missing key functionality

4. Integration and Business Value (15 points)

- **Excellent (14-15)**: Seamless integration with clear business value demonstration
- **Good (11-13)**: Good integration with adequate business value
- **Satisfactory (8-10)**: Basic integration with limited business insight
- **Needs Improvement (0-7)**: Poor integration or unclear business value

5. Innovation and Documentation (5 points)

- **Excellent (5)**: Innovative features with comprehensive documentation
- **Good (4)**: Good features with adequate documentation
- **Satisfactory (3)**: Basic features with simple documentation
- **Needs Improvement (0-2)**: Limited features or poor documentation

---

Technical Requirements

Required Technology Stack:

- **Data Processing**: Python (pandas, numpy, requests, BeautifulSoup)
- **Databases**: PostgreSQL/MySQL (structured), MongoDB (semi-structured)
- **Analytics**: scikit-learn, xgboost, statsmodels, prophet
- **Visualization**: Tableau/Power BI or Python (Plotly, Dash, Streamlit)
- **Web Framework**: Flask/Django (if custom development)

Submission Format:

1. **Complete BI Platform** - Fully functional system
2. **Technical Architecture Document** (PDF) - System design and data flow
3. **Analytics Report** (PDF) - EDA findings and model performance
4. **User Guide** (PDF) - End-user documentation with screenshots
5. **Business Case Study** (PDF) - Problem, solution, and value
6. **Code Repository** - All source code with documentation
7. **Demo Video** (10-15 minutes) - System walkthrough

Time Management Guidelines

Phase	Allocated Time	Key Focus Areas
Data Integration	60 minutes	Multi-source ETL, database design, data quality

Phase	Allocated Time	Key Focus Areas
Advanced Analytics	90 minutes	EDA, ML models, validation, optimization
Reporting System	80 minutes	Dashboard design, drill-down, ad-hoc analytics
Platform Integration	60 minutes	End-to-end workflows, BI dashboard
Advanced Features	30 minutes	Innovation, optimization, governance

---

Success Criteria

To pass this assignment, students must: - Successfully integrate structured, semi-structured, and unstructured data sources - Implement comprehensive EDA with statistical validation - Develop multiple ML models with proper validation and comparison - Create functional reporting system with drill-down and ad-hoc capabilities - Demonstrate end-to-end integration from data to insights - Provide clear business value and actionable recommendations

Bonus Points Available: - Real-time data processing implementation (+5 points) - Advanced ML techniques (ensemble methods, deep learning) (+5 points) - Superior user experience and design (+5 points) - Industry-specific innovations and optimizations (+5 points) - Cloud deployment and scalability features (+5 points)

Additional Resources

Module 2 Integration Resources:

- Multi-database connection management
- ETL best practices and error handling
- Data quality frameworks and validation
- NoSQL and SQL integration patterns

Module 3 Analytics Resources:

- Advanced statistical analysis techniques
- ML model selection and validation
- Feature engineering best practices
- Time series analysis and forecasting

Module 4 Reporting Resources:

- Dashboard design principles
- Interactive visualization techniques
- Self-service analytics implementation
- Report automation and distribution

Cross-Module Integration:

- Data pipeline orchestration tools

- Real-time analytics architectures
- Business intelligence platform design
- Analytics governance frameworks

Business Impact Measurement

Key Performance Indicators:

- **Data Quality**: Completeness, accuracy, consistency metrics
- **Analytics Performance**: Model accuracy, prediction reliability, insight relevance
- **User Adoption**: Report usage, self-service analytics engagement
- **Business Value**: Decision-making improvement, cost savings, revenue impact
- **System Performance**: Response times, scalability, reliability

Success Metrics by Scenario:

Customer Intelligence: Customer retention improvement, personalization effectiveness **Supply Chain**: Cost reduction, inventory optimization, delivery performance **Financial Risk**: Risk assignment accuracy, compliance efficiency, fraud detection

This combined assignment demonstrates the synergistic relationship between data handling, advanced analytics, and reporting, preparing students for real-world business intelligence implementations.