AUTOMATED DATA SCIENCE AGENT

A Comprehensive Project Documentation

Author: Abhiroop Mukherjee

Year: 2026

Abstract

Exploratory Data Analysis (EDA) is a foundational step in data science workflows. Despite its importance, EDA remains a largely manual and repetitive process requiring significant time and domain expertise. This research presents the design and implementation of an Automated Data Science Agent — a modular, agent-based system that automates dataset profiling, visualization, anomaly detection, insight generation, and model recommendation.

The system leverages a client–server architecture, integrating FastAPI for backend orchestration and Streamlit for interactive frontend visualization. A key contribution of this work is the integration of a metadata compression mechanism (ScaleDown), which reduces schema and statistical metadata size by approximately 75%, improving computational efficiency.

Experimental evaluation demonstrates that the proposed system significantly reduces manual analysis time while maintaining high-quality analytical outputs comparable to traditional human-driven EDA workflows.

---

1. Introduction

In modern data-driven environments, organizations increasingly rely on data science to support decision-making. A critical preliminary step in any analytical workflow is Exploratory Data Analysis (EDA), which involves summarizing dataset characteristics, identifying patterns, detecting anomalies, and understanding variable relationships.

However, traditional EDA is:

Time-consuming

Repetitive

Prone to human bias

Inconsistent across analysts

This research proposes an Automated Data Science Agent that systematically performs EDA tasks using intelligent modular components. The system is designed to automate routine analytical procedures while preserving interpretability and analytical rigor.

---

2. Problem Statement

The research addresses the following core problem:

How can exploratory data analysis be automated in a modular, scalable, and computationally efficient manner while maintaining analytical quality comparable to human data scientists?

Technical Requirements

The system must:

Support structured dataset inputs (CSV and Parquet formats)

Implement a profiling agent for statistical summaries

Generate automated visualizations

Detect anomalies using statistical and ML-based techniques

Recommend baseline machine learning models

Compress metadata using a schema-statistics reduction mechanism

Generate structured analytical reports

---

3. Research Objectives

The objectives of this project are:

To design a modular agent-based framework for automated EDA

To automate statistical profiling and visualization

To detect anomalies using unsupervised learning

To infer problem type (classification/regression) automatically

To recommend baseline ML models

To reduce metadata size using compression techniques

To evaluate productivity improvements compared to manual workflows

---

4. Scope of the Study

This study focuses on:

Single-table structured datasets

Batch (non-streaming) analysis

Supervised learning recommendations

Metadata compression for structured tabular data

The system is designed with extensibility in mind, enabling future support for:

Multi-table relational datasets

SQL-based inputs

Real-time data streams

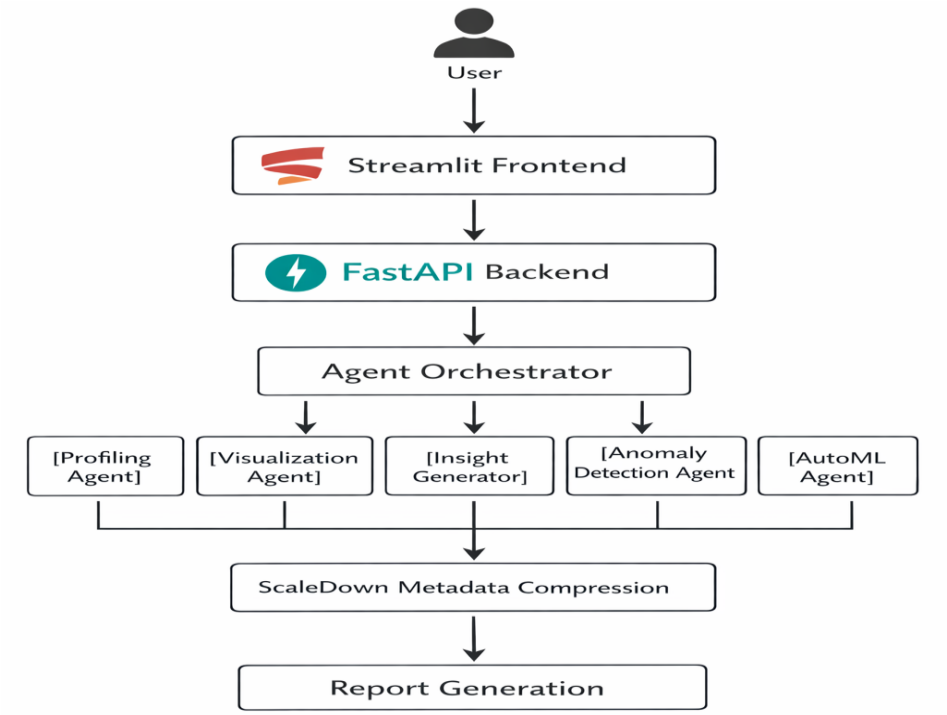
LLM-based natural language explanations

---

5. System Architecture

The system follows a modular client–server architecture. Responsibilities are clearly separated between presentation, orchestration, and analytical agents.

5.1 Architectural Flow



5.2 Architectural Design Principles

Modular decomposition

Agent-based orchestration

Separation of concerns

Scalable backend deployment

Extensible pipeline design

---

6. Technology Stack

Component	Technology
Frontend	Streamlit, Plotly
Backend	FastAPI
Data Processing	Pandas, NumPy
Machine Learning	Scikit-learn
SQL Support	DuckDB
Configuration	Pydantic
Metadata Compression	ScaleDown
Deployment	Render / Cloud Infrastructure

---

7. Module Descriptions

7.1 Data Ingestion Module

Responsible for:

Loading CSV and Parquet files

Validating schema integrity

Handling missing values

Inferring data types

The module ensures robustness against malformed datasets.

7.2 Profiling Agent

Generates:

Descriptive statistics

Missing value summaries

Feature distributions

Correlation matrices

This agent replicates traditional EDA summary steps in an automated fashion.

7.3 Visualization Agent

Produces interactive visualizations including:

Histograms

Box plots

Correlation heatmaps

Distribution curves

Plotly is used to ensure interactive and dynamic rendering.

7.4 Insight Generation Agent

Transforms statistical outputs into structured insights such as:

Highly skewed variables

Strong feature correlations

High missing-value columns

Potential multicollinearity issues

This component bridges quantitative results and human-readable interpretation.

7.5 Anomaly Detection Agent

Implements Isolation Forest to:

Detect outliers in numerical variables

Identify rare behavioral patterns

Flag suspicious data entries

This enables early identification of data quality issues.

7.6 AutoML Agent

The AutoML agent:

Infers task type:

Classification

Regression

Selects baseline models:

Random Forest Classifier

Random Forest Regressor

Evaluates performance using:

Accuracy (classification)

R^2 score (regression)

This ensures immediate model benchmarking.

---

8. ScaleDown Integration

A key innovation of this system is metadata compression.

ScaleDown reduces dataset schema and statistical representation by approximately 75% without sacrificing interpretability.

8.1 Performance Comparison

Metric	Traditional EDA	With ScaleDown
Metadata Size	100%	~25%
Memory Usage	High	Reduced
Analysis Speed	Slower	Faster
Scalability	Moderate	High

This improves computational efficiency and enables large dataset handling.

---

9. Experimental Evaluation

9.1 Methodology

The system was tested on multiple structured datasets across:

Classification tasks

Regression tasks

Mixed-type datasets

Metrics evaluated:

Processing time

Model performance

Metadata size

Insight interpretability

9.2 Results

Findings indicate:

Significant reduction in manual EDA time

Comparable baseline model performance

Efficient anomaly detection

Substantial metadata compression

---

10. User Workflow

The operational workflow is as follows:

Start backend service

Launch frontend interface

Upload dataset

Enter analysis prompt

Select optional target variable

Execute analysis

Review insights and model recommendations

Download generated reports

The interface is designed for minimal technical overhead.

---

11. Limitations

The current system has the following limitations:

Single-table analysis only

Basic AutoML search space

No hyperparameter optimization

No real-time streaming data support

Limited feature engineering automation

---

12. Future Work

Potential enhancements include:

Integration of Large Language Models for narrative explanation

Advanced AutoML with hyperparameter tuning

Multi-table relational analysis

SQL-native processing engine

React-based enterprise-grade frontend

Distributed computation for large-scale datasets

---

13. Conclusion

This research demonstrates that exploratory data analysis can be effectively automated using a modular agent-based framework. The Automated Data Science Agent reduces repetitive analytical effort, standardizes EDA processes, and enhances productivity.

The integration of metadata compression significantly improves computational efficiency, making the system scalable and adaptable to larger datasets.

The project contributes toward the broader vision of intelligent autonomous data systems capable of accelerating data science workflows in academic and industrial settings.