10Alytics

Financial Data Pipeline Optimization Program

A Data Engineering Internship project

10Alytics

Introduction

Welcome to **Gold FinTech**, where data drives our decision-making and empowers our clients to achieve their financial goals. As a leader in financial analytics and services, we are committed to delivering accurate, timely, and actionable financial information. To maintain our competitive edge and enhance our data infrastructure, we are embarking on the Financial Data Pipeline Optimization Program over the next two months. This program is essential to ensure the reliability and quality of our data, which underpins our entire operation.

10Alytics

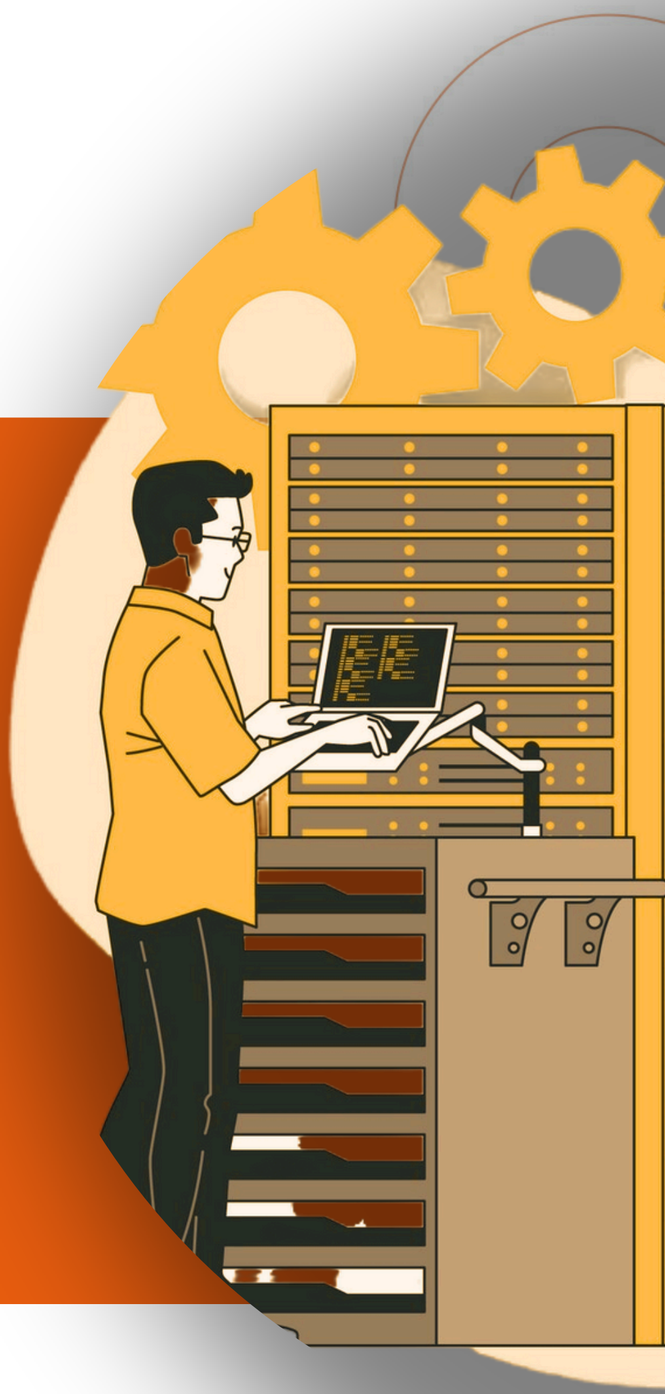
Our Company

Gold Fintech is at the forefront of financial analytics, providing cutting-edge data solutions to a diverse range of clients, from individual investors to large financial institutions. Our expertise lies in transforming raw data into insightful analytics that drive informed investment decisions. We pride ourselves on our robust data infrastructure, which supports our mission to deliver real-time, high-quality financial information. Our commitment to innovation and excellence makes us a trusted partner in the financial industry.

10Alytics

Business Problem

In today's fast-paced financial markets, timely and accurate data is crucial. However, managing vast amounts of financial data presents significant challenges. The key issues we face include the efficient extraction and processing of real-time and historical data, maintaining high data quality standards, and leveraging cloud technologies for scalability and advanced analytics. Addressing these challenges is vital for improving our service offerings and maintaining client trust.



Project Overview

10Alytics

10Alytics

Financial Market Data Project

1

Objective

Develop a pipeline to analyze financial market data to improve our data-driven insights.

Key Tasks

- Data Extraction: Extract historical and real-time data from financial APIs such as Alpha Vantage and Yahoo Finance. This will enable us to have a comprehensive dataset for analysis.
- Data Transformation: Use Apache Spark to perform data transformations and calculations. This step ensures that the data is in a usable format for analysis and reporting.
- Data Loading: Load the processed data into a relational database such as PostgreSQL. This centralized data repository will support various analytical and reporting needs.
- Pipeline Automation: Automate the pipeline to run at regular intervals using Apache Airflow. This ensures continuous data updates and reduces manual intervention.

10Alytics

End-to-End Azure Project

2

Objective

Build a data pipeline to ingest, store, transform, and load data using Azure services.

Key Tasks

- Data Ingestion: Use Azure Data Factory to ingest data from various sources into Azure Blob Storage. This ensures that data is securely transferred and stored.
- Data Storage: Organize and store raw data securely in Azure Blob Storage. This provides a scalable solution for handling large volumes of data.
- Data Transformation: Set up Azure Databricks for data cleaning and transformation. This allows for efficient and scalable data processing.
- Data Storage: Store transformed data in Azure Data Lake Storage (ADLS). This centralizes the data for further analysis and processing.
- Data Loading: Load processed data into Azure SQL Database or Synapse SQL Pools for analysis. This enables advanced querying and reporting capabilities.
- Pipeline Automation: Schedule and automate data pipelines using Azure Data Factory. This ensures that data processing occurs seamlessly and on schedule.

10Alytics

Data Quality Monitoring System

3

Objective

Develop a system to monitor and ensure the quality of data being ingested and processed.

Key Tasks

- Identify Data Quality Dimensions: Determine key data quality dimensions such as completeness, consistency, and accuracy. This helps in setting the foundation for quality assessment.
- Establish Thresholds: Set acceptable thresholds for each data quality metric. These thresholds will serve as benchmarks for our data quality checks.
- Data Quality Checks: Integrate data quality checks within the existing data pipeline using Python scripts. This step is crucial for real-time validation and error detection.
- Alert Configuration: Configure Airflow to trigger alerts via email, Slack, or PagerDuty when data quality issues are detected. This ensures that any issues are promptly addressed to maintain data integrity.

10Alytics

Deliverables

1

- Functional ETL pipeline
- Documentation of the ETL process
- PostgreSQL database with processed financial data
- Automated Airflow DAG for regular pipeline execution

2

- Configured Azure Blob Storage and Data Factory
- Data processing pipelines in Azure
- Analytical workflows in Azure Databricks
- Processed data stored in Azure Data Lake Storage and Azure SQL Database or Synapse SQL Pools
- Automated pipeline schedules in Azure Data Factory
- Documentation of the Azure data pipeline architecture

3

- Defined data quality metrics and thresholds
- Python scripts for data quality checks
- Integration of data quality checks in the pipeline
- Alerting system configured in Airflow

10Alytics

Milestones and Timeline

10Alytics

Weeks 1-2: Financial Market Data Engineering Project

Week 1:

- Project kickoff and onboarding.
- Begin data extraction from financial APIs such as Alpha Vantage and Yahoo Finance.
- Set up and configure Apache Spark for data transformations.
- Load initial datasets into PostgreSQL.

Week 2:

- Implement data transformations using Apache Spark.
- Automate the ETL pipeline using Apache Airflow.
- Test and refine the automated ETL pipeline.
- Prepare documentation of the ETL process.
- Review and finalize the Financial Market Data Engineering Project.

10Alytics

Weeks 3-4: End-to-End Azure Data Engineering Project

Week 3:

- Set up Azure Blob Storage and Data Factory for data ingestion.
- Organize and store raw data securely in Azure Blob Storage.
- Set up Azure Databricks for data cleaning and transformation.
- Implement data transformation processes in Azure Databricks.

Week 4:

- Store transformed data in Azure Data Lake Storage (ADLS).
- Load processed data into Azure SQL Database or Synapse SQL Pools.
- Schedule and automate data pipelines using Azure Data Factory.
- Test and refine the automated Azure data pipeline.
- Prepare documentation of the Azure data pipeline architecture.
- Review and finalize the End-to-End Azure Data Engineering Project.

10Alytics

Weeks 5-7: Data Quality Monitoring System

1. Define data quality dimensions and acceptable thresholds.
2. Develop initial data quality checks in Python.
3. Integrate data quality checks within the existing data pipeline.
4. Test and refine data quality checks.
5. Configure alerting mechanisms in Airflow for data quality issues.

Week 8: Report Presentation

- Prepare final project documentation and presentations.
- Conduct a review of the implemented systems.

10Alytics

Conclusion

The Financial Data Pipeline Optimization Program is critical to our strategic goals of enhancing our data infrastructure, improving data quality, and leveraging cloud technologies for better analytics. Your contributions to these initiatives will be invaluable in maintaining our position as a leader in financial analytics and services. We look forward to the innovative solutions and insights you will bring to these projects. Welcome aboard, and let's make this a productive and impactful endeavor!