
## Team Details

a. **Team name:** Zolo Hallucinators
b. **Team leader name:** Aravind S
c. **Problem Statement:** Signal Extraction from Market & News Data

## Brief about the idea

We have created an end-to-end solution that:

- Regularly ingest **market & news data** with an automated pipeline
- Maintain **historical data** in a structured database for reference
- Perform **transformations & feature generation** to create a **time-synced dataset**
- Run **ML models** to predict **tomorrow's price** & analyze **market sentiment**
- Provide **actionable signals**: short-term (tomorrow) & general trends
- Include **explainability** to trace **what influenced each prediction**
- Built a **backtest engine** to measure **strategy performance** & scoring

**Opportunities:** How different is it from any of the other existing ideas?
USP of the proposed solution?

**Our Solution is has the following USPs, and the ability to break this and provide developers with an intermediate view of data:**

- **Unified pipeline**: Combines **market data + news sentiment** in one end-to-end workflow, unlike many tools that focus on only one data type
- **Time-synced dataset**: Historical + live data **aligned for ML models**, enabling more accurate predictions
- **Explainability built-in**: Can **trace exactly what influenced a prediction**, unlike black-box models in standard platforms
- **Backtesting engine**: Evaluate **strategy performance & rewards** systematically, not just raw predictions
- **Customizable features & models**: Flexibility to **add new indicators, news sources, or ML strategies**
- **Actionable signals**: Provides **both short-term (tomorrow) and general trend signals**, bridging the gap between analytics and trading decisions
- **Data-driven decision support**: Goes beyond dashboards—turns insights into **predictive guidance**

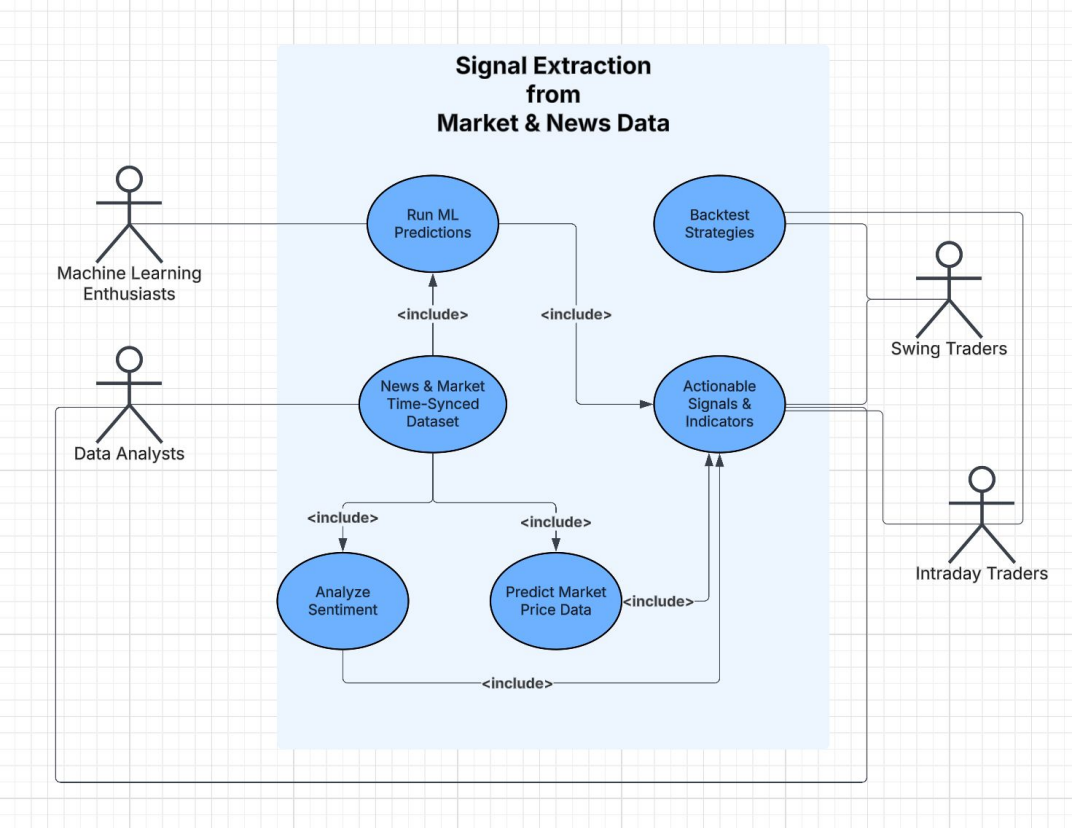
## **Opportunities:** How will it be able to solve the problem?

- **Unifies multiple data sources** (market + news) into a **single time-synced dataset** for consistent analysis
- Enables **ML-driven predictions** and sentiment scoring on a consolidated view
- Supports **strategy experimentation**: try different models to identify **winning trading strategies**
- Fills a **gap in the market**: currently, no dataset exists with this level of integration and flexibility for testing models
- **Upcoming Feature**: can plug in a **Hugging Face model link**, tune hyperparameters, and test directly on this dataset

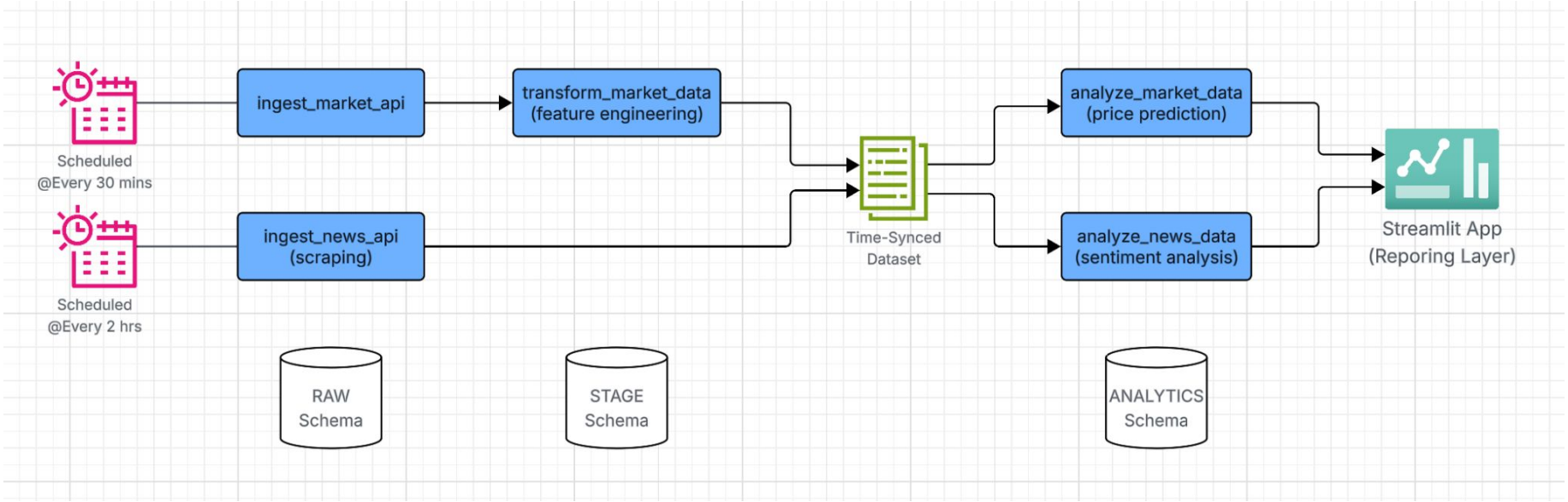
# List of features offered by the solution

- **Automated Data Ingestion** – market & news, scheduled regularly
- **Historical Data Storage** – all your past data in one place
- **Time-Synced Dataset** – aligns market & news for ML-ready analysis
- **Feature Engineering** – generate indicators, signals, sentiment scores
- **ML Predictions** – tomorrow's price & trend forecasting
- **Sentiment Analysis** – gauge market mood from news
- **Explainability** – see exactly **what drove each prediction**
- **Backtesting Engine** – test strategies, measure rewards & performance
- **Actionable Signals** – short-term & general trend guidance

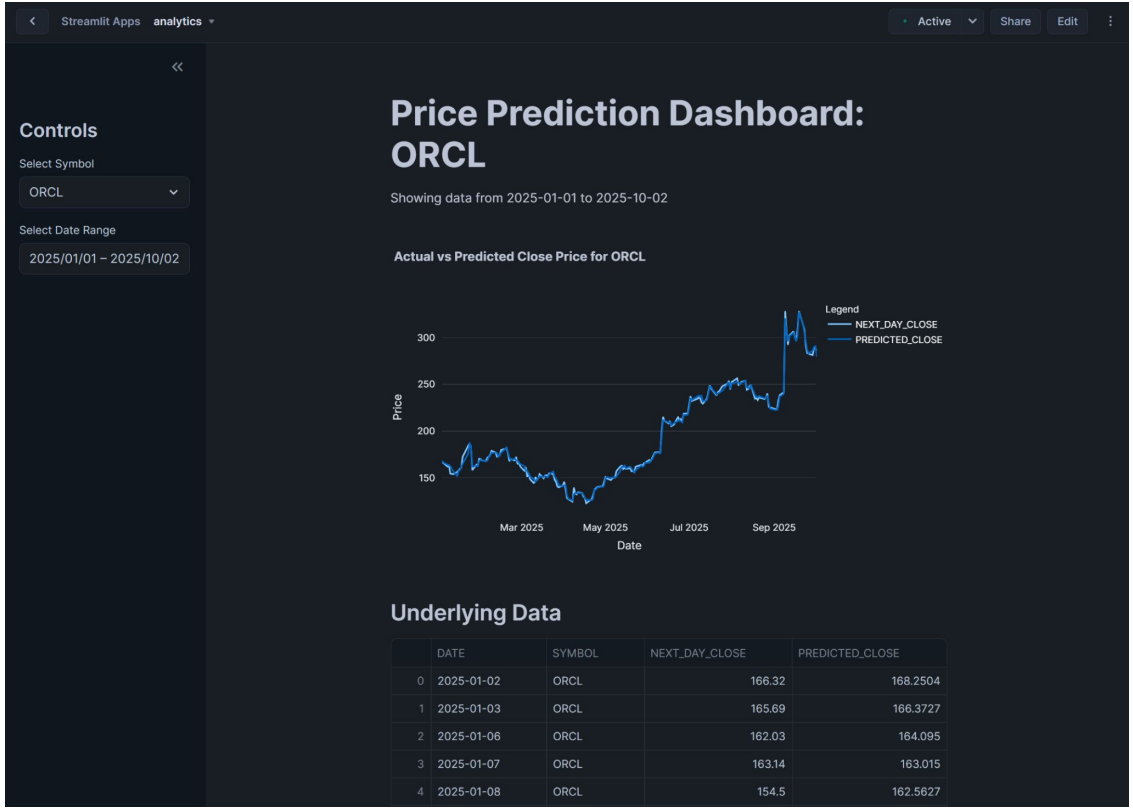
# Use-case Diagram



# Process flow diagram (inc. architecture)



# Wireframes/Mock Diagrams of the proposed solution



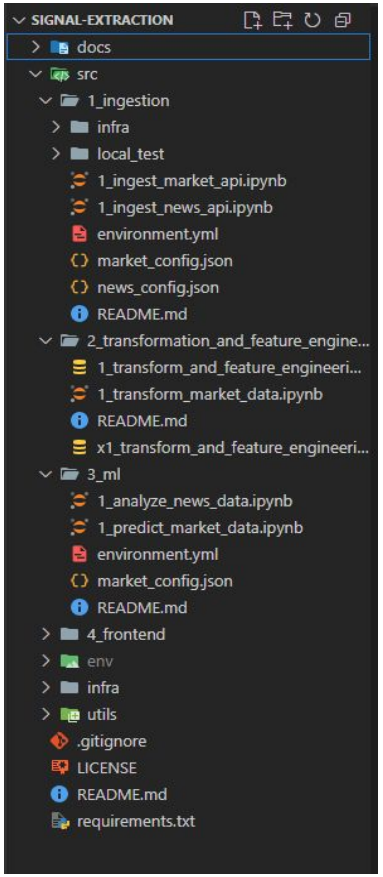
# Technologies used in the solution

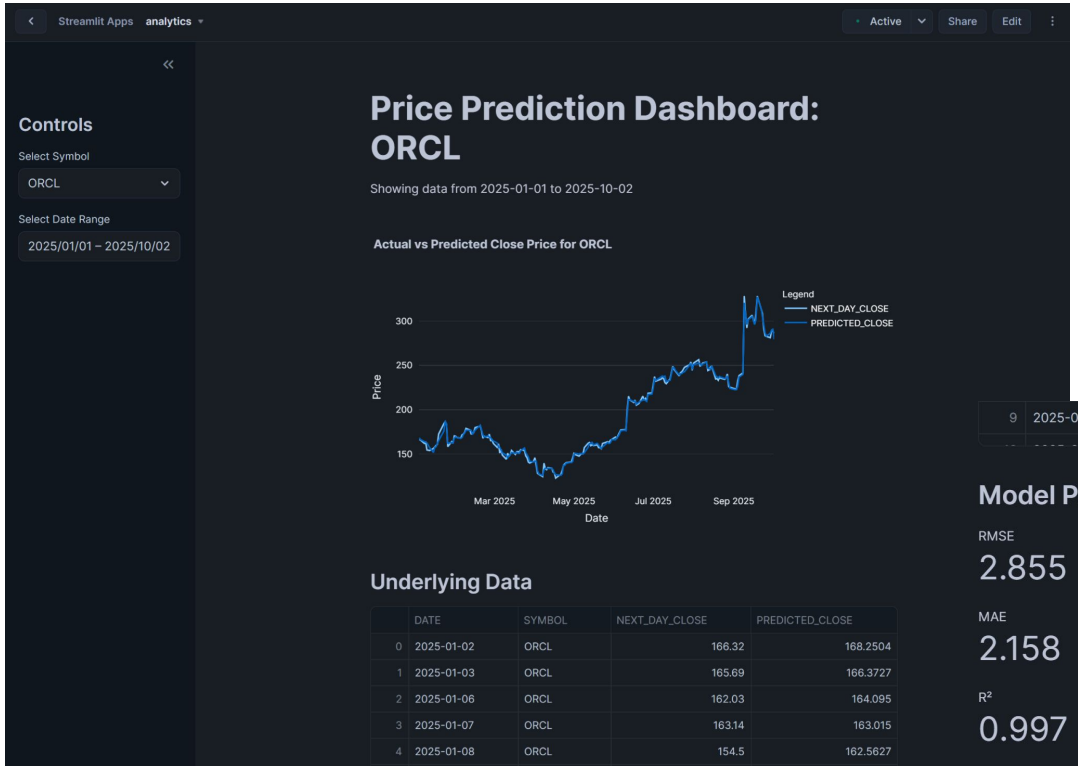
1. Ingestion & Transformation
   - Technologies: Python, APIs (Alpha Vantage, News API)
   - Components: Snowflake Tables, Medallion Arch., Snopipe, stages
2. Data Storage
   - Technologies: SQL, SQL Jobs
   - Storage: Snowflake Databases and Tables.
3. Machine Learning
   - Technologies: Python, ML Libraries, XGBoost
   - In-built functions: Snowpark ML, Cortex Search, Dataframes, Python UDFs
4. Explainability & Backtesting
   - Python, NLP Libraries, Statistics
5. Frontend/Dashboard/Reporting
   - Streamlit Application integrated in Snowflake

# Estimated implementation cost

| Title | Comment | Importance | Cost |
|---|---|---|---|
| Snowflake Business Critical Edition | For the News Ingestion Pipeline, the unique number of sites/domains that we have to scrape is in the 200s. This is only possible within this edition to have more number of network policies. | High | 4$/credit<br>~200 credits/month.<br>Totalling: 800$/month |
| newsapi.org Business Edition | In the free tier, we can only get articles up to a month with 24 hr delay, which is not enough to create high quality datasets for our ML models that need at least 3-5 years of data. | High | 449$/month |
| alphavantage.co | Free is at 25 API requests per day.<br>Monthly plans for premium membership offers 75 API requests per minute | Low | 50$/month |
| - | - | **Total** | **1299$/month** |

# Snapshots of the prototype




| | | | |
|---|---|---|---|
| 9 | 2025-01-16 | ORCL | 161.03 |

## Model Performance Metrics

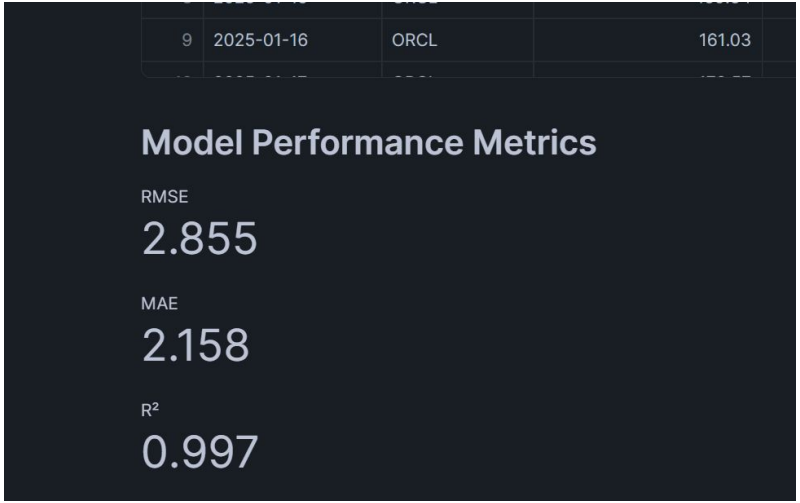
RMSE
2.855

MAE
2.158

$R^2$
0.997

# Prototype Performance report/Benchmarking

- Major focus was on building the ingestion and transformation pipeline, creating the batch processed time-synced dataset.
- Model Used: XGBRegressor
- Currently the ML model is overfitting for our case, but with more research we can integrate multiple models and better features to get better scores.



# Future Development

- **Clickable Explainability** – click any data point to see full **lineage of high-impact features**
- **Hugging Face Integration** – plug in your HF model link and test directly on our **time-synced dataset**
- **Strategy & Model Backtesting** – try different ML models & trading strategies to see **which performs best**
- **End-to-End Experimentation** – build, test, and compare models **right from the app frontend**

## Provide links to your:

**GitHub Public Repository:**
https://github.com/Zolo-Hallucinators/Signal-Extraction

**Demo Video Link:**
https://drive.google.com/file/d/1iWiBB3lU3H3SMDZ82SmHFFqlPTmzbApU/view?usp=sharing

**Final Product Link (draft):**
https://app.snowflake.com/us-east-1/lac70367/#/streamlit-apps/SIGNAL_EXTRACTION_DB.UTILS.AINREU5NXYDJBG2Y?ref=snowsight_shared

**Acknowledgement:**
I would like to sincerely thank you for this opportunity. Working on this problem statement was highly engaging, and using Snowflake for the first time was an eye-opening experience. Coming from a Databricks background, I was impressed by the platform's capabilities, which are extensive and powerful. Overall, this project has been a great learning experience, offering valuable takeaways and insights that I will carry forward.


YOURSTORY
PRESENTS
HELLO, GCC
THE DEV PREMIER
(LEAGUE);
CO-PRESENTED BY
snowflake
INNOVATION PARTNER
H2S

# Thank you