Big Data Project Proposal

1. Project Title

Final Project Proposal - MarketScope: AI-Powered Industry & Segment Intelligence Platform

2. Project Team

Yash Khavnekar
Shushil Girish
Riya Mate

3. Project Overview

MarketScope is a low-cost, AI-powered market intelligence platform designed to replace expensive and static market reports from vendors like Statista and IBISWorld. It fuses structured Snowflake data, real-time product prices and reviews, and marketing strategy literature to deliver insights on top U.S. industries and product segments. Users can explore pricing tiers, analyze TAM/SAM/SOM metrics, and receive strategy recommendations—without needing external consultants.

4. Problem Statement

Traditional market research is costly, static, and consultant-dependent, creating barriers for startups, small businesses, and independent analysts. There is a pressing need for scalable, automated tools that provide actionable insights without the premium price tag or expert interpretation.

5. Objectives

- Display macro-level outlook for top 5 U.S. industries
- Drill down into top 5 segments per industry
- Classify product segments by price tier: Cheap, Average, Luxury
- Compute TAM/SAM/SOM for each tier and segment
- Deliver book-informed strategy recommendations
- Enable real-time follow-up queries via conversational agents

6. Dataset(s)

- Snowflake Marketplace: Structured industry and segment data
- Web Scraping: Real-time product reviews and prices from Amazon, Target, Walmart

• Strategy Texts: Philip Kotler's Marketing Management, STP, 4Ps frameworks
• Estimated Volume: ~10k-100k records per segment
• Update Frequency: Weekly (Snowflake); Real-time (Web)

7. Technology Stack

• Ingestion & Orchestration: Apache Airflow, GitHub Actions
• Processing & Storage: Snowflake, AWS S3, Pinecone
• LLM Agents: LangGraph + Multi-Agent Chat Protocol
• Backend/API: FastAPI
• Frontend: Streamlit
• Modeling: Hugging Face Transformers
• Deployment: Google Cloud Run, Docker

8. Architecture Diagram

[Insert Architecture Diagram (e.g., Mingrammer, draw.io)]
Structure should show:
User → Streamlit → FastAPI → LangGraph Agents → Snowflake + Scraped Data → Pinecone for semantic search
↳ Repo Reader, Auto Doc AI, Strategy Frameworks Injection

9. Implementation Plan & Timeline

Week 1: Dataset collection, Industry and Segment list design
Week 2: Airflow DAGs for Snowflake ingestion and web scraping
Week 3: Price-tier classifier and TAM/SAM/SOM estimator implementation
Week 4: LangGraph agent development (industry summary, segmentation, review analyzer)
Week 5: Frontend with Streamlit: selection menus, visual panels, strategy Q&A
Week 6: Final deployment, testing, and documentation

10. Challenges & Risk Mitigation

• Data Inconsistencies: Validate scraped data before transformation
• Latency in Agent Responses: Optimize LangGraph execution paths
• Complex NLP Pipelines: Use modular agent design and fallback mechanisms
• Market Size Estimation Accuracy: Combine volume, price, and trend data to triangulate estimates

11. Expected Outcomes

• Fully automated intelligence platform for market research
• Live interactive dashboards for industry/segment/price analysis
• Book-informed strategy recommendations per product category

- PDF export of reports with TAM/SAM/SOM visuals
- Open-source repo with reproducible pipelines

12. References

- Philip Kotler, Marketing Management
- Snowflake Marketplace Documentation
- Amazon, Target, Walmart Public Product Pages
- LangGraph and Hugging Face Transformer APIs
- IBISWorld and Statista (for benchmark comparison)