NEXUS Project Contribution Audit Log

Project: NEXUS - Knowledge Graph-Driven Data Catalog with Unified LLM GraphRAG

Course: DAMG7374 17611 ST: LLM w/ Knowledge Graph DB

Term: Fall 2025

Institution: Northeastern University, College of Engineering Audit

Period: September 15, 2025 – December 11, 2025

Team Members

Member	Role	Primary Responsibilities
Pranav Deepak Kharat	Lead Developer & Architect	System architecture, core implementation, integration, evaluation
Venkat Akash Varun Pemmaraju	Data Engineer (Databricks)	Databricks infrastructure, synthetic data, API configuration
Shreeanant Bharadwaj	Documentation Lead	Presentation design, report compilation, documentation

---

Contribution Summary by Member

Pranav Deepak Kharat — 80% Overall Contribution

Category	Components Developed
Data Infrastructure	Snowflake account setup, Olist dataset ingestion (13 tables, 1.4M rows), schema design (OLIST_SALES, OLIST_MARKETING, OLIST_ANALYTICS)
Knowledge Graph	Neo4j schema design (4-layer structure), node/relationship definitions, 393+ nodes, 996+ relationships
GraphRAG System	Smart GraphRAG engine, Learned GraphRAG (XGBoost), Unified LLM Router, LangChain Text-to-Cypher (38 few-shot examples)
Cross-Platform Integration	Snowflake-Databricks federation, cross-source SANTOS algorithm adaptation, privacypreserving metadata matching

Category	Components Developed
Lineage & Governance	Automated lineage extraction (100% F1), SHACL-inspired validator (10 constraint shapes)
Evaluation	60-question benchmark creation, baseline implementations, statistical validation (McNemar tests)
Vector Search	Milvus integration, embedding pipeline (all-MiniLM-L6-v2), hybrid retrieval scoring
Demo Interface	Gradio 10-tab demo implementation

Venkat Akash Varun Pemmaraju — 10% Overall Contribution

Category	Components Developed
Databricks Setup	Unity Catalog configuration, workspace provisioning
Synthetic Data	Generated sales_transactions (150 rows), customer_feedback (100 rows) with sensitivity classifications
API Configuration	Databricks API token generation, authentication setup
Code Execution	Ran Databricks-side extraction scripts, validated connectivity

Shreeanant Bharadwaj — 10% Overall Contribution

Category	Components Developed
Presentation	Mid-term and final presentation slide design
Report	LaTeX/Pandoc report formatting, document compilation
Documentation	README formatting, visual asset organization

Detailed Timeline & Activity Log

Phase 1: Project Initialization (Sept 15 – Sept 30, 2025)

Date	Member	Activity	Deliverable
Sept 15	Pranav Kharat	Project scope definition, SANTOS paper review	Research notes, project proposal

Sept 16	Pranav Kharat	Snowflake account provisioning	TRAINING_DB database created
Sept 17-19	Pranav Kharat	Olist dataset acquisition and preprocessing	9 CSV files cleaned and validated
Sept 20-22	Pranav Kharat	Snowflake data ingestion pipeline	13 tables loaded (1.4M total rows)
Sept 23-24	Pranav Kharat	Schema design and organization	3 schemas: OLIST_SALES, OLIST_MARKETING, OLIST_ANALYTICS
Sept 25-27	Pranav Kharat	Neo4j local setup, initial ontology design	Core node types defined
Sept 28	Venkat Pemmaraju	Databricks workspace creation	Unity Catalog provisioned
Sept 29-30	Pranav Kharat	Knowledge graph schema finalization	9 relationship types defined

Phase 2: Core Development (Oct 1 – Oct 31, 2025)

Date	Member	Activity	Deliverable
Oct 1-3	Pranav Kharat	Neo4j KG builder implementation	kg_builder.py, olist_kg_builder.py
Oct 4-5	Pranav Kharat	OlistData and OlistColumn node creation	95 Layer 3 nodes
Oct 6-8	Pranav Kharat	Sample data loading (Customer, Order, Product)	298 Layer 2 nodes
Oct 9-11	Pranav Kharat	SANTOS algorithm research and adaptation	Score calculation formula defined
Date	Member	Activity	Deliverable
Oct 12-14	Pranav Kharat	Within-Snowflake duplicate detection	OLIST_DUPLICATE edges (100% accuracy)
Oct 15-17	Pranav Kharat	Milvus vector database setup	Docker compose, collection schema
Oct 18-19	Pranav Kharat	Embedding pipeline (sentence-transformers)	vector_indexer.py (384-dim embeddings)

Oct 20-22	Pranav Kharat	Smart GraphRAG engine v1	Initial 43% accuracy
Oct 23-25	Pranav Kharat	Hybrid ranking formula optimization	80/20 semantic/structural weighting
Oct 26	Venkat Pemmaraju	Synthetic data schema design	sales_transactions, customer_feedback schemas
Oct 27-28	Venkat Pemmaraju	Synthetic data generation	250 rows with sensitivity tags
Oct 29	Venkat Pemmaraju	Databricks data upload	Tables registered in Unity Catalog
Oct 30	Venkat Pemmaraju	API token generation	Databricks authentication configured
Oct 31	Shreeanant Bharadwaj	Mid-term presentation outline	Slide structure defined

Phase 3: Integration & Advanced Features (Nov 1 – Nov 20, 2025)

Date	Member	Activity	Deliverable
Nov 1-3	Pranav Kharat	Databricks metadata extractor	databricks_metadata_extractor.py
Nov 4-5	Pranav Kharat	Federated KG builder	FederatedTable, FederatedColumn nodes (42 nodes)
Nov 6-8	Pranav Kharat	Cross-source SANTOS implementation	cross_source_duplicate_detector.py
Nov 9-10	Pranav Kharat	SIMILAR_TO relationship creation	16 cross-source matches detected

Date	Member	Activity	Deliverable
Nov 11-12	Pranav Kharat	Snowflake lineage extractor	QUERY_HISTORY parsing, DERIVES_FROM edges
Nov 13	Pranav Kharat	Lineage graph builder	6 lineage edges (100% F1)

Nov 14-15	Pranav Kharat	SHACL-inspired governance validator	10 constraint shapes, 3 severity levels
Nov 16-17	Pranav Kharat	LangChain Text-to-Cypher integration	4 query-type-specific prompts
Nov 18	Pranav Kharat	Few-shot example curation	38 Cypher examples across query types
Nov 19	Pranav Kharat	Unified LLM GraphRAG router	Intent classification (100% accuracy)
Nov 20	Venkat Pemmaraju	Databricks code execution validation	Connectivity verified end-to-end

Phase 4: Evaluation & Optimization (Nov 21 – Nov 30, 2025)

Date	Member	Activity	Deliverable
Nov 21-22	Pranav Kharat	Benchmark question creation	60 questions, 4 categories
Nov 23	Pranav Kharat	Ground truth labeling	Expert annotations for all queries
Nov 24-25	Pranav Kharat	Baseline implementations	Keyword, Embeddings-Only, Graph-Only
Nov 26	Pranav Kharat	XGBoost learned routing	train_route_classifier.py, model artifacts
Nov 27	Pranav Kharat	Comparative evaluation framework	run_comparative_evaluation.py
Nov 28	Pranav Kharat	Statistical significance testing	McNemar tests, p-value calculations
Nov 29	Pranav Kharat	Explainable GraphRAG	WHY explanations for cross-source matches
Nov 30	Shreeanant Bharadwaj	Mid-term presentation slides	10-slide deck completed

Phase 5: Demo & Documentation (Dec 1 – Dec 11, 2025)

Date	Member	Activity	Deliverable
Dec 1-2	Pranav Kharat	Gradio demo interface	10-tab professional UI
Dec 3	Pranav Kharat	Tab implementations (Search, Lineage, Compare)	Tabs 1-3 functional
Dec 4	Pranav Kharat	Tab implementations (Duplicates, Performance)	Tabs 4-5 functional
Dec 5	Pranav Kharat	Tab implementations (System, Governance, Federation)	Tabs 6-8 functional
Dec 6	Pranav Kharat	Tab implementations (Features, About)	Tabs 9-10 functional

Dec 7	Pranav Kharat	End-to-end system testing	All 6 RQs validated
Dec 8	Shreeanant Bharadwaj	Report structure and LaTeX setup	Pandoc template configured
Dec 9	Shreeanant Bharadwaj	Report content compilation	Sections 1-10 drafted
Dec 10	Shreeanant Bharadwaj	Report finalization	Appendices, references added
Dec 11	All Members	Final review and submission	Complete project package

Code Contribution Summary

File/Module	Lines	Author	Description
demo_gradio.py	1,873	Pranav Kharat	10-tab demo interface
src/graphrag/few_shot_examples.py	1,450	Pranav Kharat	38 Cypher examples with documentation
data/evaluation/benchmark_questions.json	1,200	Pranav Kharat	60-question evaluation benchmark

File/Module	Lines	Author	Description
src/knowledge_graph/olist_kg_builder.py	890	Pranav Kharat	Olist-specific KG construction logic
notebooks/exploration.ipynb	850	Pranav Kharat	Data exploration and prototyping
src/graphrag/unified_llm_graphrag.py	820	Pranav Kharat	Master query router
src/graphrag/smart_graphrag_engine.py	780	Pranav Kharat	Rule-based hybrid retrieval
src/graphrag/langchain_graphrag.py	720	Pranav Kharat	Text-to-Cypher pipeline
src/federation/cross_source_duplicate_detector.py	680	Pranav Kharat	SANTOS adaptation
src/graphrag/learned_graphrag_engine.py	580	Pranav Kharat	XGBoost routing baseline
src/connectors/snowflake_connector.py	540	Pranav Kharat	Snowflake metadata extraction
src/knowledge_graph/kg_builder.py	520	Pranav Kharat	Core KG construction
src/federation/federated_kg_builder.py	490	Pranav Kharat	Cross-platform KG nodes
src/graphrag/vector_indexer.py	450	Pranav Kharat	Milvus embedding pipeline
README.md	450	Pranav Kharat	Project documentation
tests/test_graphrag.py	420	Pranav Kharat	GraphRAG unit tests
scripts/run_comparative_evaluation.py	420	Pranav Kharat	Benchmark execution
src/graphrag/explainable_graphrag.py	380	Pranav Kharat	WHY explanation generator
tests/test_kg_builder.py	380	Pranav Kharat	KG builder unit tests
src/utils/cypher_templates.py	369	Pranav Kharat	Cypher query templates
src/governance/shacl_validator.py	360	Pranav Kharat	Constraint validation
scripts/load_json_to_kg.py	350	Pranav Kharat	JSON data loader
tests/test_federation.py	340	Pranav Kharat	Federation unit tests
src/extractors/metadata_extractor.py	340	Pranav Kharat	Schema parsing utilities
src/lineage/snowflake_lineage_extractor.py	320	Pranav Kharat	Query history parsing

File/Module	Lines	Author	Description
src/federation/databricks_metadata_extractor.py	310	Venkat Akash Pemmaraju	Databricks API integration
src/utils/helpers.py	290	Pranav Kharat	Utility functions
src/lineage/lineage_graph_builder.py	290	Pranav Kharat	DERIVES_FROM creation
scripts/load_olist_to_kg.py	280	Pranav Kharat	KG population script
src/utils/embeddings.py	260	Pranav Kharat	Embedding utilities
scripts/olist_uploader.py	240	Pranav Kharat	Data loading utility
scripts/train_route_classifier.py	220	Pranav Kharat	XGBoost training
config/settings.py	180	Pranav Kharat	Configuration management
main.py	180	Pranav Kharat	ETL orchestration
docker-compose.yaml	120	Pranav Kharat	Container orchestration
requirements.txt	80	Pranav Kharat	Python dependencies
Total	**18,422**		

Research Contribution Mapping

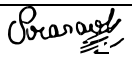
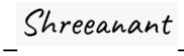
Research Question	Primary Contributor	Supporting Contributor
RQ1: GraphRAG vs Embeddings	Pranav Kharat	Shreeanant Bharadwaj
RQ2: Automated Lineage	Pranav Kharat	Venkat Pemmaraju
RQ3: SHACL Governance	Pranav Kharat	Shreeanant Bharadwaj
RQ4: Rules vs ML	Pranav Kharat	Venkat Pemmaraju
RQ5: Hybrid vs Pure Neural	Pranav Kharat	Shreeanant Bharadwaj
RQ6: Cross-Source SANTOS	Pranav Kharat	Venkat Pemmaraju

Artifacts & Deliverables

Deliverable	Owner	Status
GitHub Repository (18,422 LOC, 715 deletions)	Pranav Kharat	✅ Complete
Neo4j Knowledge Graph (393 nodes)	Pranav Kharat	✅ Complete
Milvus Vector Index (15 embeddings)	Pranav Kharat	✅ Complete
60-Question Benchmark	Pranav Kharat	✅ Complete
Evaluation Results JSON	Pranav Kharat	✅ Complete
Databricks Unity Catalog (2 tables)	Venkat Pemmaraju	✅ Complete
Final Presentation (PDF)	Shreeanant Bharadwaj	✅ Complete
Technical Report (40 pages)	Shreeanant Bharadwaj	✅ Complete

Certification

We, the undersigned, certify that this audit log accurately represents the contributions of each team member to the NEXUS project during Fall 2025.

Member	Signature	Date
Pranav Deepak Kharat		December 11, 2025
Venkat Akash Varun Pemmaraju		December 11, 2025
Shreeanant Bharadwaj		December 11, 2025

Document generated: December 11, 2025

Version: 1.0