Research Proposal

Cross-Functional Frameworks for Petabyte-Scale Analytics: Design, Optimization, and Cost Modeling

Background:

Today, organizations generate petabytes of data in areas like retail, finance, healthcare, and IoT. Processing and analyzing this data requires frameworks that bring together data engineering, distributed computing, and cost optimization strategies. Most current solutions focus on separate parts, such as storage or computing, instead of taking a complete approach to build scalable systems, improve performance, and model costs in hybrid environments (cloud plus on-prem).

Problem Statement

Petabyte-scale analytics faces challenges such as:

- Fragmented architectures lacking interoperability between data engineering, ML, and BI teams.
- High operational costs due to inefficient resource allocation.
- Performance bottlenecks in distributed query execution and ETL pipelines.
- Limited predictive cost modeling for multi-cloud and hybrid deployments.

Objectives

1. Design a cross-functional framework that unifies data engineering, analytics, and ML workflows for petabyte-scale datasets.
2. Develop optimization strategies for query execution, storage tiering, and compute resource allocation.
3. Create predictive cost models for hybrid and multi-cloud deployments using AI-driven simulations.
4. Benchmark performance and cost trade-offs across different architectures (e.g., Spark, Presto, Snowflake, BigQuery).

Proposed Approach

- Architecture Design: Define modular components for ingestion, transformation, and analytics using distributed systems (Apache Spark, Presto, Dask).

- Optimization Techniques:
  - Adaptive query optimization using ML.
  - Intelligent caching and tiered storage strategies.

- Cost Modeling:
  - Build AI-driven models to predict cost under varying workloads.
  - Simulate scenarios for hybrid cloud vs. on-prem deployments.

- Evaluation:
  - Use real-world datasets (e.g., retail transaction logs, IoT streams).
  - Metrics: Query latency, throughput, cost per TB processed.

Methodology

- A reference architecture for petabyte-scale analytics.
- An optimization toolkit for performance tuning.
- A predictive cost modeling framework for enterprise decision-making.
- Comparative analysis of cost-performance trade-offs across platforms.

Conclusion

This research will help enterprises scale analytics efficiently, lower costs, and improve teamwork between data engineering, analytics, and ML teams. It responds to the rising demand for frameworks that work together amid massive data growth and hybrid cloud use.