CSP554: Big Data Technologies

Project Draft

Project Title: Clickstream Funnel Analytics on a NewSQL Warehouse

Team Members:

- **Manushi Patel** - mpatel188@hawk.illinoistech.edu
- **Atharv Rathore** - arathore6@hawk.illinoistech.edu
- **Samarth Rajput** - srajput4@hawk.illinoistech.edu
- **John Isaiah** - ijohn@hawk.illinoistech.edu

---

1. Project Overview

This project aims to design and implement a **real-time clickstream analytics pipeline** using **NewSQL technologies** to demonstrate fast, scalable analytics on semi-structured web event data. Using **Wikimedia pageview logs**, the system will produce metrics such as funnel conversion, user activity, and drop-off patterns in real time.

The proposed stack leverages **Kafka** for ingestion, **Spark Structured Streaming** for processing, **Snowflake** as the analytical warehouse, and **Grafana/Streamlit** for dashboard visualization. The project demonstrates the end-to-end flow of a modern cloud-based big data pipeline that unifies streaming and warehousing capabilities.

---

2. Literature Review

2.1 NewSQL and Analytical Warehousing

Modern analytical workloads require horizontally scalable systems that combine transactional and analytical capabilities. **NewSQL** platforms such as Snowflake, Google BigQuery, and AWS Redshift provide these hybrid features through distributed storage, parallel execution, and SQL compatibility.

2.2 Comparison: Snowflake vs Redshift vs BigQuery

- **Snowflake** provides a multi-cluster shared data architecture with independent scaling of compute and storage, making it cost-effective for variable workloads. It stores semi-structured data using the **VARIANT** type and supports schema-on-read queries.

- **Google BigQuery** is serverless and optimized for high concurrency. It uses **columnar storage** and **Dremel execution** to perform near real-time analytics. However, it lacks the fine-grained compute control of Snowflake.
- **AWS Redshift** provides cluster-based control with local SSD caching and materialized views for performance tuning. While highly performant, Redshift can be less flexible for unstructured JSON ingestion compared to Snowflake.

Studies such as Chen et al. (2022) have shown that Snowflake outperforms traditional OLAP systems when handling semi-structured data at scale, while BigQuery excels in ad-hoc query latency. Together, these insights inform our technology selection.

2.3 Data Profiling and Exploration

The team will perform **data profiling** to characterize the Wikimedia clickstream dataset. Using **Spark DataFrames** and **Snowflake queries**, we will compute:

- Minimum, maximum, mean, and standard deviation for numeric attributes.
- Null count and frequency distribution for categorical fields (e.g., referrer type, page ID).
- Value distributions across time-based and session-based groupings.

Profiling results will be visualized through charts and tables in the project report.

---

3. Technical Approach

3.1 Data Ingestion

A **Python producer** will replay Wikimedia pageview JSON data as an event stream into **Kafka** hosted on AWS MSK. Data will be stored temporarily in Snowflake's **VARIANT** column for transformation.

3.2 Transformation and Modeling

Using **Spark Structured Streaming on EMR**, the team will perform sessionization and window-based aggregations.

3.3 Visualization

The **Grafana/Streamlit dashboard** will display:

- Funnel conversion by navigation path
- Active sessions per minute
- Top drop-offs and P95 latency

---

4. Project Milestones and Owners

Milestone	Description	Owner(s)	Due Date
1. Data Profiling & Ingestion	Set up Kafka producer; profile Wikimedia pageviews; validate schema quality	**Manushi Patel, Samarth Rajput**	Nov 16, 2025
2. Warehouse Setup	Create Snowflake warehouse and stage raw VARIANT data	**Atharv Rathore, John Isaiah**	Nov 20, 2025
3. Transform & Sessionization	Implement Spark Structured Streaming on EMR for session aggregation	**Samarth Rajput, Atharv Rathore**	Nov 27, 2025
4. Dashboard Development	Design Grafana/Streamlit dashboard for visualization	**Manushi Patel, John Isaiah**	Dec 3, 2025
5. Benchmarking & Final Report	Performance evaluation and report completion	**All Members**	Dec 8, 2025

---

5. Expected Deliverables

- End-to-end pipeline with Kafka + Spark + Snowflake.
- Interactive funnel dashboard for session analytics.
- Final write-up with references and performance summary.

---

6. References

1. Stonebraker, M. (2012). "NewSQL: An Alternative to NoSQL and Old SQL for New OLTP/OLAP Problems."
2. Armbrust et al. (2015). "Structured Streaming: A Declarative API for Real-Time Applications in Apache Spark."
3. Chen, J., et al. *"Stream Analytics on Modern Warehouses."* IEEE Data Engineering Bulletin, 2022.
4. Gartner. *"Emerging Data Warehouse Architectures for Real-Time Analytics."* Gartner Report, 2023.
5. Wikimedia Foundation. *"Wikimedia Pageview Clickstream Dataset."* 2023. https://dumps.wikimedia.org/other/clickstream/
6. Snowflake Inc. *"VARIANT Data Type and Semi-Structured Data Guide."* 2023.
7. Google Cloud. *"BigQuery Architecture and Performance Optimization."* 2023.
8. Amazon Web Services. *"Amazon Redshift Architecture Overview."* AWS Whitepaper, 2023.