PySpark Clickstream Analytics - Walkthrough

Overview

This project implements a scalable data analytics pipeline for website clickstream data using PySpark. It simulates a modern data engineering stack with Bronze, Silver, and Gold layers using Delta Lake.

Prerequisites

- Python 3.12
- Java 11
- PySpark 3.5.0
- Delta Spark 3.0.0

Pipeline Architecture

1. **Data Generation**: Synthetic clickstream data (JSON) simulating user navigation.
2. **Bronze Layer**: Raw data ingestion into Delta tables.
3. **Silver Layer**: Sessionization logic (30-minute inactivity timeout) to group events into sessions.
4. **Gold Layer**: Aggregated metrics (session duration, page views) and funnel analysis.

Execution Steps

1. Setup Environment

We created a virtual environment and installed dependencies:

```
python3.12 -m venv venv_312
./venv_312/bin/pip install pyspark==3.5.0 delta-spark==3.0.0 pandas
matplotlib setuptools
```

2. Generate Data

Generated 5000 synthetic events:

```
./venv_312/bin/python generate_data.py
```

3. Run Pipeline

Processed data through Bronze -> Silver -> Gold layers:

```
./venv_312/bin/python pipeline.py
```

Output Sample (Gold Layer):

```
+----------------+-------------------+-------------------+----------+--------------
|unique_session_id|           end_time|
start_time|page_views|duration_seconds|add_to_cart| ...

+----------------+-------------------+-------------------+----------+--------------
|       user_083_6|2025-11-20 06:37:14|2025-11-20 06:22:51|
```

```
15|            863|          3| ...
+----------------+-----------------+-----------------+----------+--------------
```

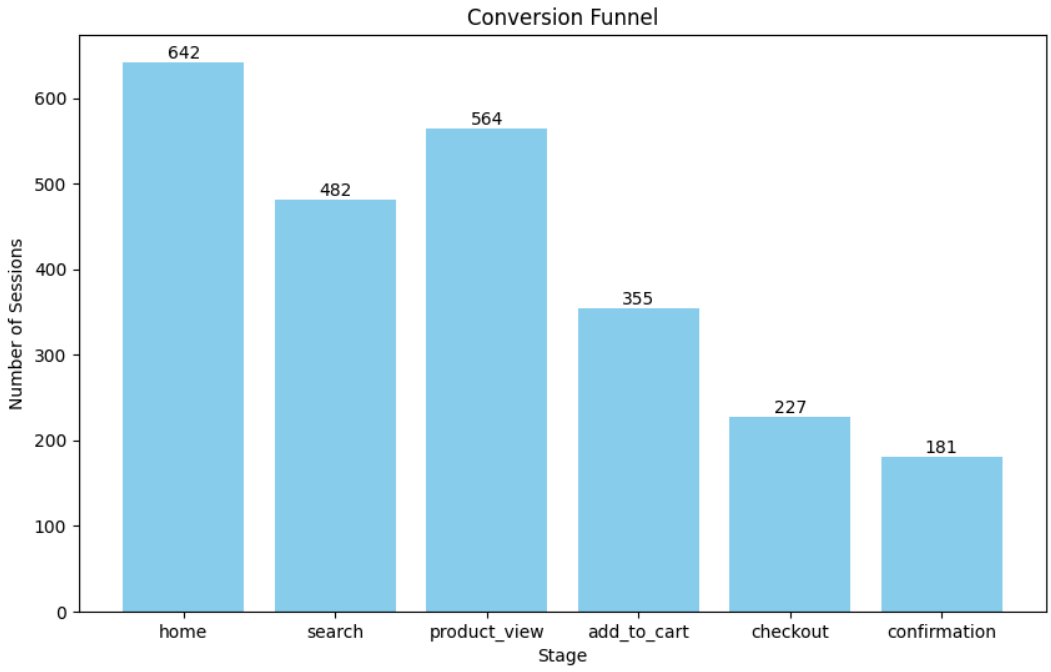
4. Visualization

Generated funnel and session duration charts:

```
./venv_312/bin/python visualize.py
```

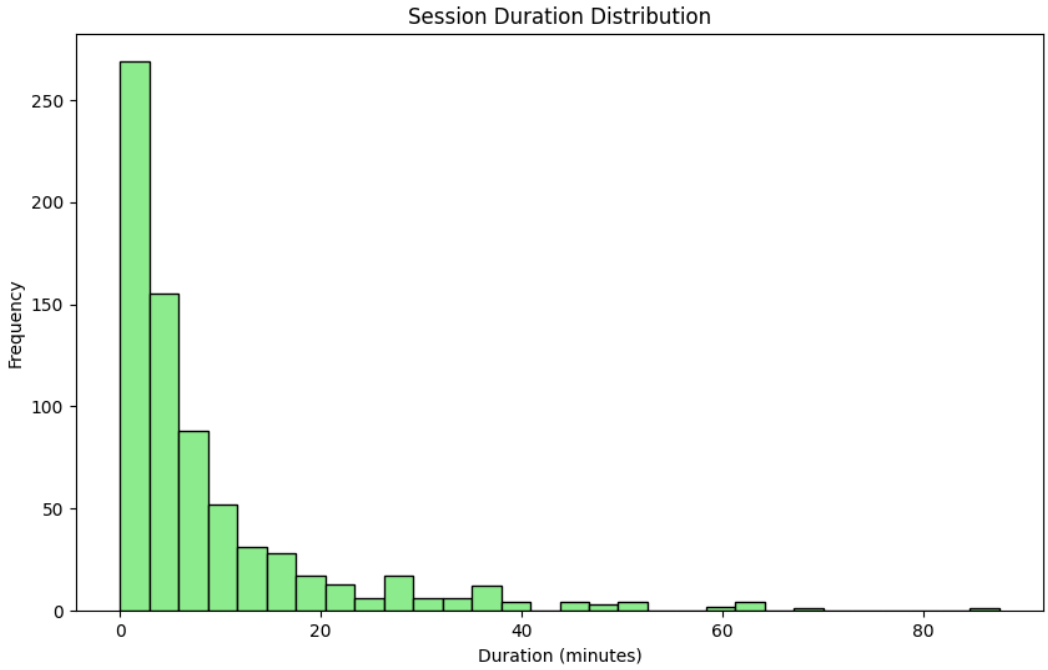
Results

Conversion Funnel



Conversion Funnel

Session Duration



Session Duration

Key Metrics

- **Sessionization**: Successfully identified sessions based on 30-minute gaps.
- **Funnel**: Tracked conversion from Home -> Search -> Product -> Cart -> Checkout -> Confirmation.