Uber NYC Pickup Analysis using Apache Spark

# Executive Summary

This project demonstrates a comprehensive big data analytics pipeline processing 1,028,136 Uber trips from September 2014 in New York City using Apache Spark. The analysis reveals critical insights into temporal patterns, geographic hotspots, and operational efficiency that can inform business strategy and resource allocation.

# 1. Project Overview

## 1.1 Dataset Specifications

Source: Uber raw data for September 2014 (Kaggle)  
  
Records: 1,028,136 individual Uber pickups  
  
Key Fields:  
  
Date/Time: Timestamp of trip  
  
Lat/Lon: GPS coordinates of pickup location  
  
Base: Uber base/affiliate code  
  
Data Quality: High integrity with proper schema enforcement

## 1.2 Technology Stack

Apache Spark 4.0.1 (Distributed processing engine)  
  
PySpark (Python API for Spark)  
  
Pandas & Matplotlib (Visualization and analysis)  
  
Local cluster deployment for scalable processing

# 2. Technical Implementation

## 2.1 Data Ingestion & Schema Enforcement

Schema definition for data quality:  
- Date/Time: StringType  
- Lat: DoubleType  
- Lon: DoubleType  
- Base: StringType  
Achievement: Successfully loaded 1M+ records with zero data type conversion errors.

## 2.2 Data Transformation Pipeline

Timestamp Conversion: M/d/yyyy H:mm:ss format parsing  
  
Feature Extraction:  
- Hour of day (0-23)  
- Day of week (1-7)  
- Date formatting  
  
Data Quality: 0 null values after datetime conversion

## 2.3 Geographic Data Cleaning

NYC Coordinate Bounds: Latitude (40.4-41.0), Longitude (-74.5, -73.5)  
  
Data Quality Result: 99.7% records within valid NYC bounds  
  
Records Cleaned: 1,025,450 of 1,028,136

# 3. Core Analysis & Findings

## 3.1 Temporal Analysis

Peak Hour Identification  
Rank Hour Pickups Time Period  
1 18:00 75,040 Evening Rush  
2 17:00 73,373 Evening Rush  
3 19:00 69,660 Evening Rush  
4 16:00 68,224 Late Afternoon  
5 15:00 61,219 Afternoon  
Insight: Clear evening rush hour pattern with 6 PM being the absolute peak.  
  
Weekly Demand Patterns  
Day Pickups Rank  
Tuesday 163,230 1  
Saturday 162,057 2  
Friday 160,380 3  
Thursday 153,276 4  
Monday 137,288 5  
Wednesday 135,373 6  
Sunday 116,532 7  
Insight: Mid-week (Tue-Thu) shows consistently high demand, with Sunday being the slowest day.

## 3.2 Geographic Hotspot Analysis

Top 5 Pickup Locations  
Latitude Longitude Pickup Density  
40.76 -73.98 43,848  
40.74 -73.99 41,732  
40.76 -73.97 41,700  
40.75 -73.99 40,174  
40.73 -74.00 35,203  
Key Finding: Manhattan areas around coordinates (40.76, -73.98) represent the highest density pickup zone.  
  
Geographic Distribution Statistics  
Unique Locations: 2,623 distinct coordinate pairs  
Average Density: 391 pickups per location  
Maximum Density: 43,848 pickups (top location)  
Minimum Density: 1 pickup (isolated locations)

## 3.3 Operational Analysis by Base

Base Performance Metrics  
Base Pickups Percentage Rank  
B02617 377,695 36.74% 1  
B02598 240,600 23.40% 2  
B02682 197,138 19.17% 3  
B02764 178,333 17.35% 4  
B02512 34,370 3.34% 5  
Insight: Base B02617 handles over one-third of all Uber trips, indicating superior operational capacity.

## 3.4 Hotspot Concentration Analysis

Threshold-Based Density Distribution  
Threshold Locations Coverage  
10+ pickups 876 99.5% of all pickups  
50+ pickups 446 98.5% of all pickups  
100+ pickups 309 97.5% of all pickups  
200+ pickups 228 96.4% of all pickups  
Strategic Insight: Focusing on just 228 high-density locations (8.7% of total) covers 96.4% of all pickup demand.

# 4. Key Business Insights

## 4.1 Demand Patterns

Peak Hours: 5-7 PM (evening commute) shows 210,000+ combined pickups  
  
Weekly Cycle: Tuesday peak demand likely correlates with business travel patterns  
  
Geographic Concentration: Manhattan dominates with specific high-density corridors

## 4.2 Operational Efficiency

Base Disparity: Top base (B02617) performs 10x better than lowest (B02512)  
  
Resource Allocation: 96% of demand covered by just 9% of locations

## 4.3 Strategic Opportunities

Surge Pricing: Evening hours present optimal pricing opportunities  
  
Driver Allocation: Pre-position drivers in top 200 locations during peak hours  
  
Base Optimization: Replicate B02617's successful operational model

# 5. Technical Achievements

## 5.1 Big Data Processing Capabilities

✅ Distributed Processing: Handled 1M+ records using Spark's distributed architecture  
  
✅ Data Quality: Maintained 99.7% data integrity through rigorous cleaning  
  
✅ Scalable Analytics: All analyses performed using Spark DataFrame operations  
  
✅ Performance Optimization: Proper caching and partition management

## 5.2 Analytical Depth

✅ Multi-dimensional Analysis: Temporal, geographic, and operational perspectives  
  
✅ Statistical Rigor: Comprehensive density and distribution analysis  
  
✅ Actionable Insights: Business-ready recommendations with data backing  
  
✅ Visualization Ready: Structured output for dashboard integration

# 6. Conclusion

This project successfully demonstrates enterprise-grade big data analytics using Apache Spark. The analysis provides Uber with:  
  
Evidence-based decision support for driver allocation and surge pricing  
  
Operational intelligence for base performance optimization  
  
Strategic insights for geographic expansion and resource planning  
  
Technical blueprint for scalable data processing pipelines  
  
The implementation proves that distributed computing with Spark can transform raw location data into actionable business intelligence at scale.

**Name: M.Devisri**

**Roll No: 2211CS010666 (S2-86)**