Proposal Part 1

**Team Members**: Pavan Kalyan Imadabathini

Anushya Sivanupandian

Aasritha Devi Surapaneni

**Project Name**: Geospatial Analysis Using Apache Spark

**Group Name**: Team Devils

IFT 512: Advanced Analytics Big Data/AI

Prof. Robert Rucker

October 8th, 2023

**Summary**

The proposed project represents a comprehensive geospatial analysis initiative focused on New York City's Yellow Cab taxi trip records. Spanning from January 2009 to June 2015, this extensive dataset presents a wealth of information relevant to urban planning, transportation management, and business intelligence. By harnessing advanced technologies such as Apache Spark, Hadoop Distributed File System (HDFS), Scala, and Amazon EC2 virtual machines, this project aims to unlock the latent value within this data, facilitating data-driven decision-making processes.

Geospatial analysis has emerged as a critical discipline across various domains, enabling the identification of clusters, outliers, and trends within geographic data. In this project, the spotlight is on the taxi trip records, a data source that holds significant importance in understanding urban mobility patterns and optimizing transportation services. The project's primary focus will be on data ranging from January 2009 to June 2015, effectively encompassing the entire lifecycle of NYC's taxi data and providing a comprehensive foundation for analysis.

The motivation for selecting this project task is multifaceted. Firstly, geospatial analysis is gaining increasing prominence in our data-driven society, offering insights that can drive strategic decisions. Secondly, the project's embrace of cutting-edge technologies, including Apache Spark and HDFS, positions it at the forefront of contemporary data analysis trends. Additionally, this project presents a unique opportunity for participants to gain expertise in the realms of distributed computing and spatial statistics, skillsets that are increasingly valuable in today's competitive job market.

To execute this project effectively, a robust data foundation is essential. The project will rely on the rich dataset of New York City Yellow Cab taxi trip records, covering an extensive temporal range. Furthermore, monthly datasets from 2009 to 2012 will supplement the analysis for hotspot identification. These datasets are sourced from official records, ensuring both the accuracy and relevance of the data. Rigorous data preprocessing, encompassing cleaning and transformation steps, will be undertaken to ensure that the data is primed for advanced analysis.

The project's technological backbone consists of state-of-the-art tools and platforms. Apache Spark, renowned for its capabilities in distributed data processing, provides a computational engine, enabling scalable and efficient analysis of vast geospatial datasets. The Hadoop Distributed File System (HDFS) serves as the storage backbone, offering not only the distributed storage capacity required for large datasets but also fault tolerance and accessibility. Scala, a versatile and powerful programming language, has been selected for the implementation of user-defined functions and SparkSQL queries. Finally, Amazon EC2 virtual machines provide the necessary computational resources for setting up the Spark cluster, ensuring that ample computing power is available to handle the analysis tasks effectively.

The project unfolds in two distinct phases. Phase 1 is dedicated to the implementation of essential spatial queries. This includes a range of queries such as range, range join, distance, and distance join queries. These queries will be constructed using SparkSQL and user-defined functions, allowing for efficient extraction of relevant geospatial insights from the data. Phase 2 transitions into hotspot analysis, a more advanced form of geospatial analysis. Here, the project will apply spatial statistics to identify statistically significant spatial hot spots using the Getis-Ord statistic. Importantly, this analysis will be conducted using monthly NYC taxi trip data, adding a temporal dimension to the geospatial insights.

An additional layer of structure is provided by the project's adherence to the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework. This framework guides the project through a systematic and comprehensive approach to data mining and analysis. It spans crucial steps, including business understanding, data understanding, data preparation, modeling, evaluation, and deployment, ensuring that the project is well-structured, and results driven.

In conclusion, this project represents a substantial opportunity to harness the power of geospatial analysis in enhancing decision-making processes, particularly in the realms of urban planning, transportation optimization, and business intelligence. Its potential benefits to organizations and industries are far-reaching, offering insights that can drive more informed and efficient operations. As a continuous effort, the project remains open to future expansion, which may involve incorporating additional data sources or applying the methodology to different geographic regions, further extending its impact and relevance.

1. **Why are you choosing this task?**

**Ans:**

**Relevance:** Geospatial analysis is pivotal in today's data-driven world. This project aligns with modern industry trends by utilizing big data technologies like Apache Spark and HDFS. It offers a valuable opportunity to acquire expertise in distributed computing and spatial statistics, highly sought-after skills in the job market.

**Challenging and Relevant Problem**: This task is chosen because it presents a challenging and relevant problem in the field of geospatial data analysis and spatial statistics. It involves implementing spatial hot spot analysis using Apache Spark, which is a powerful distributed computing framework. The task also requires the calculation of the Getis-Ord statistic, a well-established method for identifying statistically significant spatial clusters or hot spots.

**Practical Value**: This task offers the opportunity to apply advanced analytical techniques to real-world data, making it intellectually stimulating and practically valuable. It provides insights into urban planning, transportation management, and business intelligence, contributing to more informed decision-making.

**Rich Dataset:** The focus on New York City's Yellow Cab taxi trip records from January 2009 to June 2015 offers a wealth of data, vital for understanding urban transportation patterns.

**Structured Methodology:** Adhering to the CRISP-DM framework ensures a systematic approach, covering key steps from data understanding to deployment, enhancing the project's rigor.

**Cutting-Edge Technologies**: The project employs Apache Spark and Scala, which are well-suited for the efficient processing of large geospatial datasets. These technologies enable scalability and provide practical experience with state-of-the-art tools.

**Impact:** By extracting insights from geospatial data, the project contributes to improved decision-making, particularly in urban planning, transportation management, and business intelligence. The use of the Getis-Ord statistic adds statistical depth to the analysis, enhancing its real-world relevance and potential impact.

In conclusion, this project's selection is driven by its challenging and relevant nature, the opportunity to apply advanced techniques, the richness of the dataset, the adoption of cutting-edge technologies, and the potential for significant real-world impact.

**2. What are the organizational benefits of doing so, that is, list at least 5 questions Our project will address?**

**Ans:**

**5 questions our proposal will address:**

* How can geospatial insights from taxi trip data improve decision-making in urban planning, transportation management, and business intelligence?
* How can hotspot analysis assist in optimizing the allocation of resources, such as taxi services, to areas with higher demand, leading to increased operational efficiency?
* How can the identification of statistically significant clusters or hotspots help organizations proactively mitigate risks, such as traffic congestion or service disruptions?
* How can the utilization of advanced geospatial analysis techniques provide organizations with a competitive advantage by uncovering hidden market trends or customer behavior patterns?
* How can the project's findings be leveraged to formulate data-driven strategies that enhance service quality, customer satisfaction, and overall operational performance?

3. **What data set(s) are you proposing to use?**

**Ans:**

1. **New York City Yellow Cab Taxi Trip Records**: This dataset spans from January 2009 to June 2015 and contains a comprehensive collection of taxi trip records in New York City. It includes information about pick-up and drop-off locations, timestamps, fares, and other relevant attributes. This dataset forms the core of our analysis, enabling us to perform spatial queries and hotspot analysis.
2. **Monthly NYC Taxi Trip Datasets (2009-2012):** In addition to the main dataset, we will incorporate monthly NYC taxi trip datasets from the years 2009 to 2012. These datasets provide a spatio-temporal perspective and are used specifically for hotspot analysis using the Getis-Ord statistic. Each monthly dataset offers insights into taxi trip patterns and hotspots for that timeframe.

These datasets are sourced from official records, ensuring data accuracy and relevance to our geospatial analysis project. Data preprocessing will be conducted to clean, format, and transform the data, making it suitable for analysis.

4. **What technologies do you propose to use?**

**Ans:** The proposed project will leverage the following key technologies:

1. **Apache Spark**: Apache Spark will serve as the core computational engine for distributed data processing. Its parallel processing capabilities make it well-suited for handling large geospatial datasets efficiently.
2. **Hadoop Distributed File System (HDFS):** HDFS will be used as the distributed storage system for storing and managing geospatial data. It provides fault tolerance and scalability, essential for big data applications.
3. **Scala:** Scala, a versatile programming language, will be employed for implementing user-defined functions and SparkSQL queries. Its functional programming features make it an excellent choice for data manipulation tasks.
4. **Amazon EC2:** Amazon Elastic Compute Cloud (EC2) virtual machines will provide the necessary computational resources to set up the Spark cluster. EC2 instances can be easily configured to meet the project's computing requirements.

These technologies together create a robust and scalable framework for geospatial data analysis, enabling the project to efficiently process, analyze, and derive valuable insights from large and complex datasets.

5. **Now comment on the CRISP steps and how Our project matches them?**

**Ans:**

The CRISPDM (CrossIndustry Standard Process for Data Mining) framework consists of several welldefined steps for approaching and executing data mining projects. Let's examine how Our project matches these steps:

1. **Business Understanding**: Our project begins with a clear understanding of the business objectives. We aim to extract geospatial insights from NYC Yellow Cab taxi trip data to enhance decision-making in urban planning, transportation management, and business intelligence.
2. **Data Understanding**: We thoroughly explore and understand the dataset, which includes NYC Yellow Cab taxi trip records spanning from January 2009 to June 2015. We also incorporate monthly datasets from 2009 to 2012 for hotspot analysis, ensuring comprehensive data understanding.
3. **Data Preparation**: Data preprocessing is a critical step. We clean and transform the dataset to ensure data accuracy and readiness for analysis. This step involves handling missing values, data formatting, and filtering relevant attributes.
4. **Modeling**: The project involves two main phases. In Phase 1, we implement spatial queries using Apache Spark and user-defined functions. In Phase 2, we apply hotspot analysis, utilizing the Getis-Ord statistic to identify statistically significant spatial hot spots.
5. **Evaluation:** We evaluate the results of our geospatial analysis to ensure they align with the project's objectives. Evaluation criteria include the significance of identified hot spots and their potential impact on decision-making.
6. **Deployment:** While the primary focus is on analysis, the project's findings are intended for deployment in real-world scenarios. The insights derived can be used to inform decision-makers in urban planning, transportation, and business intelligence.

Throughout the project, we iterate on these steps as needed to ensure a systematic and results-driven approach to geospatial data analysis. The CRISP-DM framework provides a structured methodology that guides our project's execution, enhancing its rigor and effectiveness in delivering valuable insights and actionable recommendations.

**References:**

<https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf>

<https://www.oracle.com/autonomous-database/what-is-geospatial-database/>

<http://sigspatial2016.sigspatial.org/giscup2016/problem>

<https://www.tutorialspoint.com/spark_sql/spark_sql_quick_guide.htm>

<https://spark.apache.org/docs/latest/>