

MSBX5420 Team Project – Requirement Specification

Team : Team Blanca Peak

Date: 04/12/2020

Team Members :

Chaerin Lee	Chaerin.Lee@colorado.edu
Dean Duke	Dean.Duke@colorado. edu
Ethan Goldbeck	Ethan.Goldbeck@colorado.edu
Matthew Kuchar	Matthew.Kuchar@colorado.edu
Soumya Panda	Soumya.Panda@colorado.edu

Document History

Version	Date	Author	Comments
1.0	4/12/2020	Team Blanca Peak	Initial Version
1.1	4/19/2020	Team Blanca Peak	Updated the contents

Table of Contents

PURPOSE	3
DATASET	3
FUNCTIONAL REQUIREMENTS.....	6
NON FUNCTIONAL REQUIREMENTS	6
PERFORMANCE REQUIREMENTS	6
PROJECT GOALS AND TIMELINE.....	7
SYSTEM REQUIREMENT.....	8
ASSUMPTIONS AND DEPENDENCIES:	8
REFERENCES:.....	8
ABBREVIATIONS / ACRONYMS:	8

Purpose

The purpose of this document is to document the requirements for the MSBX5420 project of the team Blanca Peak. We will describe our dataset, the functional, nonfunctional, and performance requirements, as well as the overall goals and timetable of the project.

Dataset

We have selected the NYC taxi dataset (Yellow Taxi Trip Record 2019) for our team project. The New York City Taxi and Limousine Commission (TLC), created in 1971, is the agency responsible for licensing and regulating New York City's medallion (yellow) taxis, street hail livery (green) taxis, for-hire vehicles (FHV), commuter vans, and paratransit vehicles. The TLC collects trip record information for each taxi and for-hire vehicle trip completed by licensed drivers and vehicles. The yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, etc. The list of fields and the field descriptions are given below.

Field Name	Description
VendorID	A code indicating the TPEP provider that provided the record. 1= Creative Mobile Technologies LLC; 2= VeriFone Inc.
tpep_pickup_datetime	The date and time when the meter was engaged.
tpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.

RateCodeID	<p>The final rate code in effect at the end of the trip.</p> <p>1= Standard rate</p> <p>2=JFK</p> <p>3=Newark</p> <p>4=Nassau or Westchester</p> <p>5=Negotiated fare</p> <p>6=Group ride</p>
Store_and_fwd_flag	<p>This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server.</p> <p>Y= store and forward trip</p> <p>N= not a store and forward trip</p>
PULocationID	TLC Taxi Zone in which the taximeter was engaged
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged
Payment_type	<p>A numeric code signifying how the passenger paid for the trip.</p> <p>1= Credit card</p> <p>2= Cash</p> <p>3= No charge</p> <p>4= Dispute</p> <p>5= Unknown</p>

	6= Voided trip
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges. Currently, this only includes the \$0.50 and \$1 rush hour and overnight charges.
MTA_tax	\$0.50 MTA tax that is automatically triggered based on the metered rate in use.
Tip_amount	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Total amount of all tolls paid in trip.
Improvement_surcharge	\$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
Total_amount	The total amount charged to passengers. Does not include cash tips.

Functional Requirements

Process	Process Name: Ingest Data in HDFS/S3
The script should upload the Yellow Taxi 2019 dataset into either HDFS or AWS S3 system.	
Req No	Requirement
URS-1	ssh to the above system and upload the dataset from the local system.
URS-2	Ensure the dataset is saved in the system

Process	Process Name: Data Analysis
Data Analysis of the Yellow Taxi 2019 dataset	
Req No	Requirement
URS-1	Connect the saved dataset using Spark and perform data analysis such as view, count, aggregate, group etc

Process	Process Name: Data Visualization
Req No	Requirement
URS-1	Display visualization based on data analysis

Process	Process Name: Spark Machine Learning
Req No	Requirement
URS-1	Create machine learning model using regression/classification

Non Functional Requirements

Performance Requirements

Req No	Performance Requirement
URS-1	Add node to the cluster
URS-2	Measure the performance by adding the new node

Project Goals and Timeline

Goals:

- Determine busiest areas in NYC for picking up/dropping for a given period
- Describe the 'average ride' for a given period
- How has the Yellow Cab Market changed over time
- Understand the impact of the introduction/expansion of ridesharing companies (Uber, Lyft) on the Yellow Cab market in NYC
- Other potentially interesting findings about NYC taxis over the past decade
- Visualize finding

Timeline

Immediate:

Data Preparation (Cleaning, Compiling, Transforming, etc.)

By April 25th:

Designing, Developing and Testing

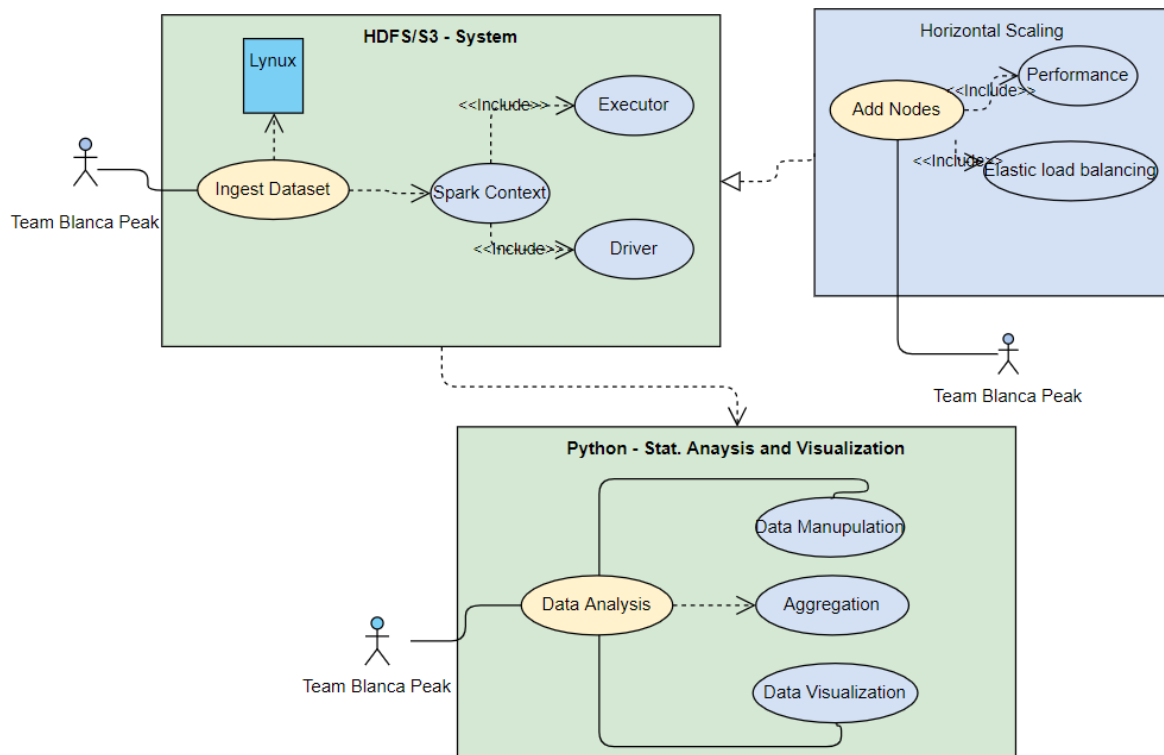
By April 28th:

Deployment

April 28th:

Presentation

System Requirement



Assumptions and Dependencies:

The system is provided with below components.

1. **HDFS/S3** : The data storage system.
2. **Spark with Python** : The system to perform the data analysis, machine learning, manipulation and visualization.
3. **Testing** : Since we have time constraints hence we go with functional/manual testing.

References:

<https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

Abbreviations / Acronyms:

Abbreviation Acronym	Description
URS	User Requirement Specification