Leeds School of Business
UNIVERSITY OF COLORADO BOULDER

MSBX5420 Team Project – Requirement Specification

Team : Team Blanca Peak **Date:** 04/12/2020

Team Members :

Kuchar, Matthew
Lee, Chaerin
Panda, Soumya
Goldbeck, Ethan
Duke, Dean

Document History

Version	Date	Author	Comments
1.0	4/12/2020	Team Blanca Peak	Initial Version

Table of Contents

Introduction

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 (FHVs), 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	The final rate code in effect at the end of the trip. 1= Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride
Store_and_fwd_flag	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. Y= store and forward trip N= not a store and forward trip
PULocationID	TLC Taxi Zone in which the taximeter was engaged
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged

Payment_type	A numeric code signifying how the passenger paid for the trip. 1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 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
URS-2	

Process	Process Name: **Data Visualization**
Req No	**Requirement**
URS-1	Display monthly revenue
URS-2	Display passenger count by by month

Non Functional Requirements

Security Requirements

Req No	Security Requirement
URS-1	Data should only be accessible to the internal team

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 overtime
- Understand the impact of the introduction/expansion of ridesharing companies (Uber, Lyft) on the Yellow Cab market in NYC
- Visualize finding

Timeline

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

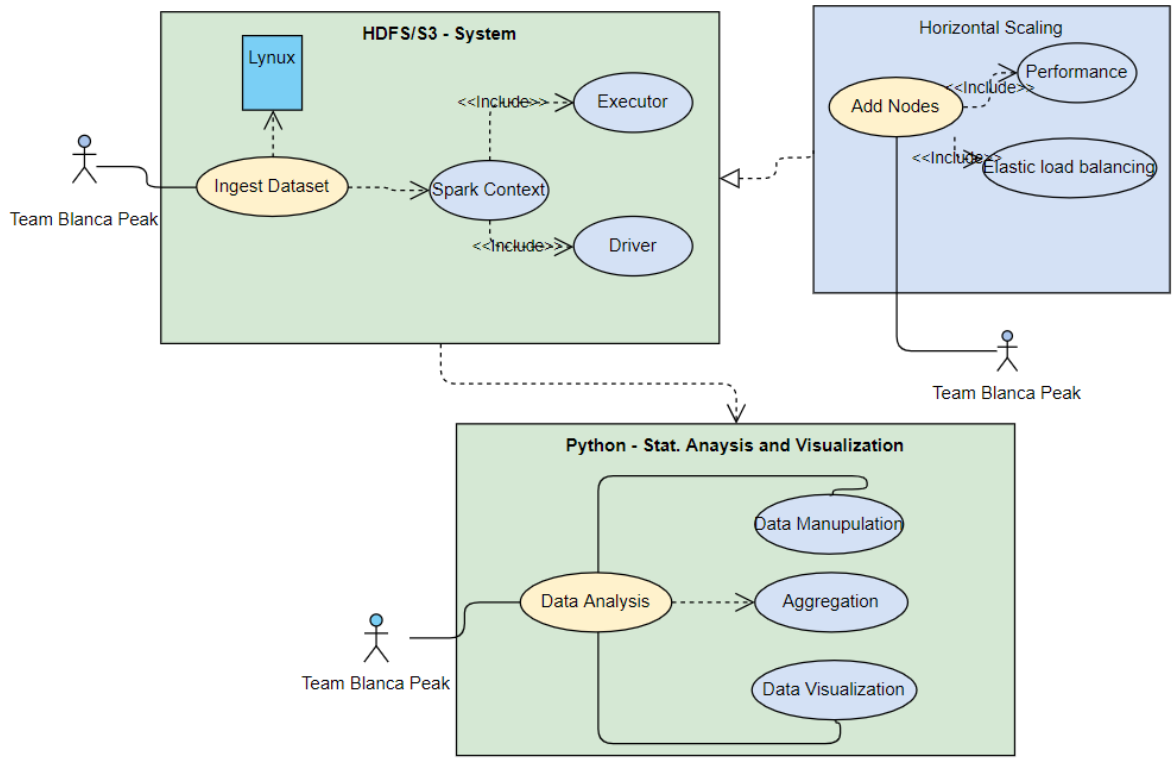
By April 25th:
Designing, Developing and Testing

By April 28th:
Deployment

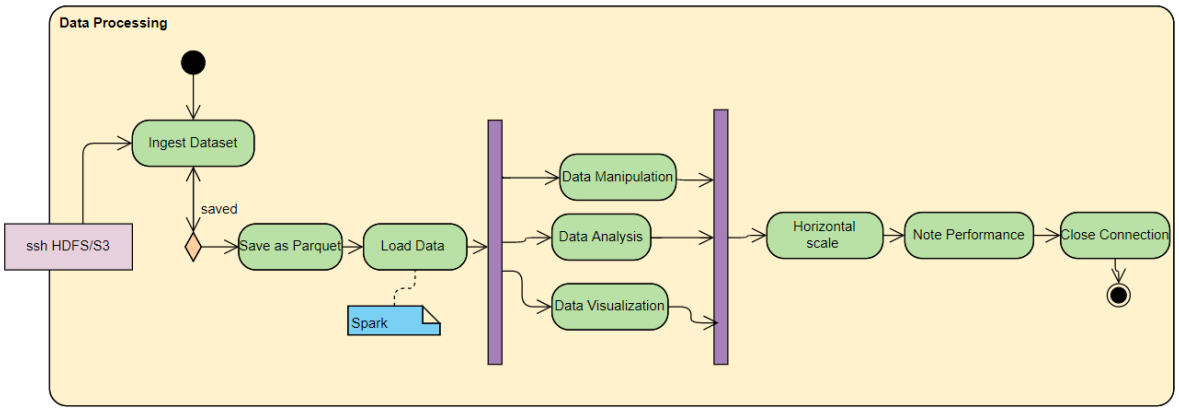
April 28th:
Presentation

System Design

Use Case Design :



Activity Design :



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, manipulation and visualization.
3. **Testing :** Unit testing, integration testing, life cycle testing to ensure the system meets the requirement specification.

References, Abbreviations/Acronyms

References:

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

Abbreviations / Acronyms:

Abbreviation Acronym	Description
URS	User Requirement Specification