



Big Data Analytics Framework for NYC Green Taxi

Muhammad Nidzam bin Maso’od

Analytics Vision

mnidzam@analytics.vision.com

# Executive Summary

### [**The Market**](https://www.thebalancesmb.com/how-to-write-a-business-plan-section-3-2947029)

Across the world has seen an explosion of growth of Big Data analytics over the last couple of years. Big Data applications and analytics is projected to grow from $5.3B in 2018 to $19.4B in 2026, attaining a CAGR of 15.49%. Big Data market worldwide includes Professional Services is projected to grow from $16.5B in 2018 to $21.3B in 2026. Due to explosion of data collection and data analytics, more company are moving to big data infrastructures such as moving their traditional database to Hadoop clustering and the usage of Spark to speed up Big Data analytics process. Due to our expertise in the field, NYC Green Taxi has request us to design and implement a Big Data Analytics framework for them.

### [**Our Competitive Advantages**](https://www.thebalancesmb.com/how-to-write-the-competitive-analysis-section-of-the-business-plan-2947025)

* While there are currently many others company offering Big Data platform, we are the only few which specialize to help our customer to build big data analytics framework from infrastructure setup till the design of analytics user interface and visualization.
* Our strategy is to emphasize the quality and competitive cost of service we provided. Help Desk services has also been setup to provide a 24x7 support to our customer to resolve any technical issues that may arise after product has been successfully deployed in our customer environment.
* Our engineers are being trained and certified with the latest Big Data technologies especially in Hadoop and Spark.
* All employees are [insured and bonded](https://www.thebalancesmb.com/why-and-how-to-get-bonded-in-canada-2948240).

### [**Financial Projections**](https://www.thebalancesmb.com/writing-the-business-plan-section-8-2947026)

Based on the calculation we have done, we projected profit of $80000 for this project.

### [**Financing Requirements**](https://www.thebalancesmb.com/starting-a-business-finding-small-business-financing-2948106)

We are seeking an operating line of $150,000 to finance this project.

Contents

[Executive Summary 2](#_Toc529614157)

[**The Market** 2](#_Toc529614158)

[**Our Competitive Advantages** 2](#_Toc529614159)

[**Financial Projections** 2](#_Toc529614160)

[**Financing Requirements** 2](#_Toc529614161)

[Introduction 4](#_Toc529614162)

[Proposed Big Data Framework 4](#_Toc529614163)

[Data Analytics pipeline 7](#_Toc529614164)

[Hadoop Hardware 8](#_Toc529614165)

[Timeline 10](#_Toc529614166)

[Proof of concept 11](#_Toc529614167)

[Run Hadoop on test machine 11](#_Toc529614168)

[Run Spark and Jupyter Notebook on test machine 14](#_Toc529614169)

[Use Spark to load NYN Green Taxi 2015 data from Hadoop 15](#_Toc529614170)

[Costing 19](#_Toc529614171)

[Summary 21](#_Toc529614172)

# Introduction

Our customer, NYC Green Taxi is embarking on a Big Data transformation plan. We have been asked to help them in their journey by designing and implementing Big Data Analytics framework for the company. In this report, we will go through the proposed framework and infrastructure needed for this project.

## Proposed Big Data Framework

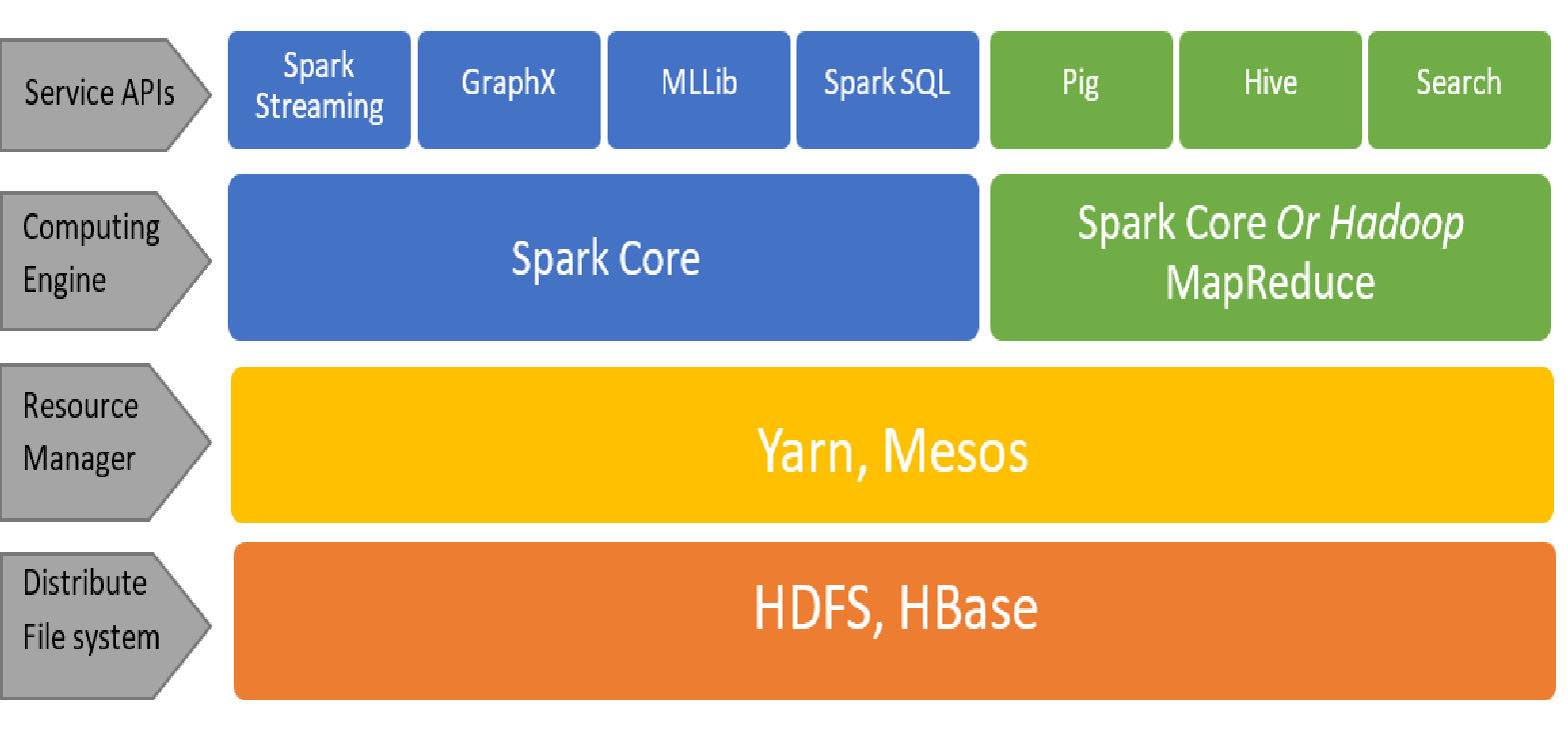


Figure 1 Proposed Big Data Framework for NYC Green Taxi

For this project, our company is proposing Apache Hadoop as the file storage system and for the computing engine we will be proposing Apache Spark. The selection

Hadoop is an open source Big Data solution. Since it is an open source solution, Hadoop’s cost per terabyte is much less than high end data warehouse database server like Teradata and Oracle’s Exadata. Hadoop’s storage costs are also substantially less than many high and mid-level storage area network (SAN) solutions. Hadoop architecture is a distributed file storage system, making it very cost effective in replacing other data storage solution. Hadoop also have some function to enable user to be able to query data from its storage.

Hadoop’s basic storage structure is a distributed file system called HDFS (Hadoop Distributed File System).

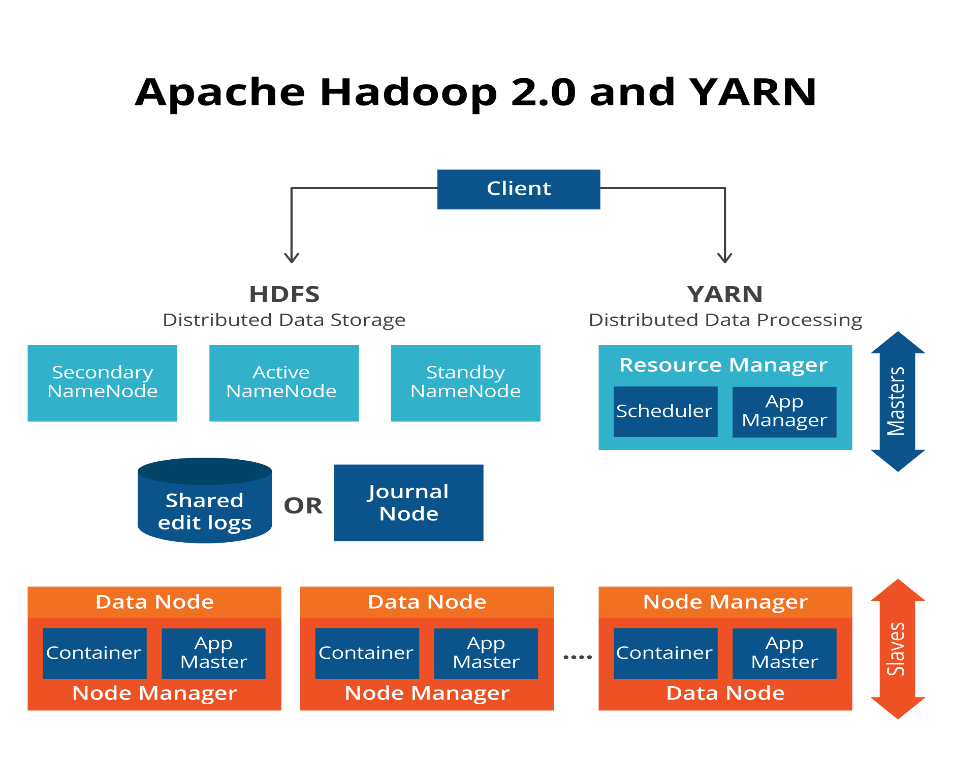


Figure 2: Hadoop Storage Structure

 It distributes data very simply but effectively.  HDFS automatically makes 3 copies of the entire file across 3 separate computer nodes of the Hadoop cluster (a node is a commodity Intel server).  While duplicating data three times does consume more space than other methodologies, Hadoop’s storage cost is so much less than standard systems, it is actually a practical, cost-effective solution.  If a node goes offline, HDFS has two other remaining copies.

A Hadoop cluster is built using racks of commodity Intel servers with each server having internal hard disk storage.  One 72 inches floor rack can hold 42 1U servers. To mitigate the risk of failing hard drive, Hadoop automatically make 3 copies of file to 3 separate computer nodes. This will provide a high reliability file storage system. In the event of disk failure, data can be restored using one of the computer nodes thus making Hadoop to have a very small risk or almost no risk to permanent data loss.

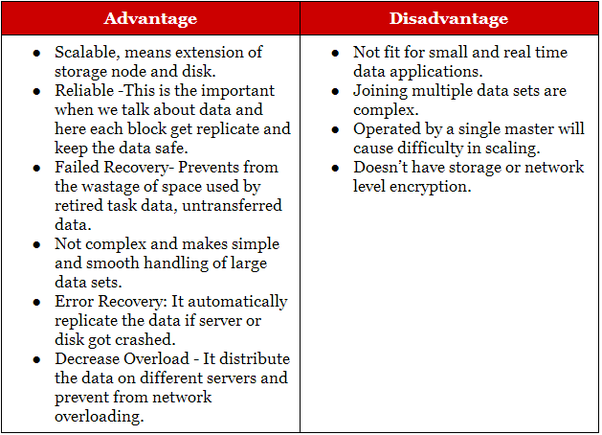


Figure 3: Advantage and disadvantages of Hadoop

Our proposed framework will be using Spark as the computing engine over Hadoop Map Reduce. Spark is a fast cluster computing system developed by the contributions of near about 250 developers from 50 companies in the UC Berkeley’s AMP Lab, for making data analytics faster and easier to write and as well to run.

Apache Spark is an open source available for free download thus no additional cost required. The software can be installed inside our proposed Hadoop infrastructure and being deployed via Hadoop Yarn. Spark uses in-memory computing and optimization of arbitrary operator graphs compared to MapReduce which use disk base engines making the querying process to be much faster. For programming language, Java, Scala and python is supported

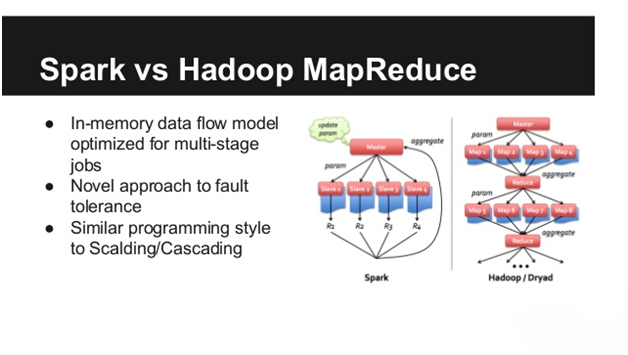


Figure 4: Spark vs Hadoop MapReduce

## Data Analytics pipeline

Our proposed data analytics pipeline can be visualized as below figure.

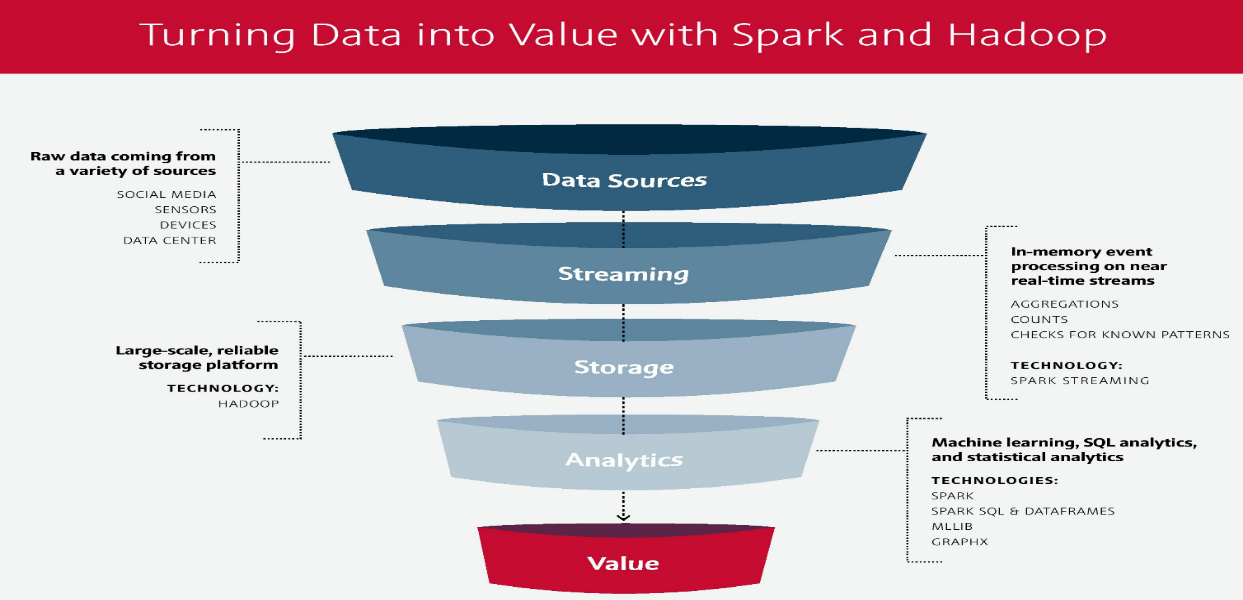


Figure 5 Data Analytics pipeline

Data from NYC Green Cab data provider vendor will be stream and stored inside Hadoop. For analytics purposes, Spark engine will be used to do calculations. Data then can be visualized using various visualization tools.

## Hadoop Hardware

For Small clusters—clusters with fewer than 20 worker nodes—do not require much for master nodes in terms of hardware. A solid baseline hardware profile for a cluster of this size is a dual quad-core 2.6 Ghz CPU, 24 GB of DDR3 RAM, dual 1 Gb Ethernet NICs, a SAS drive controller, and at least two SATA II drives in a JBOD configuration in addition to the host OS device. Clusters of up to 300 nodes fall into the mid-size category and usually benefit from an additional 24 GB of RAM for a total of 48 GB. Master nodes in large clusters should have a total of 96 GB of RAM.

For this project, we are proposing to use server from Server Direct (<https://www.serversdirect.com/>). The proposed machine specification and pricing is as below. is as below.

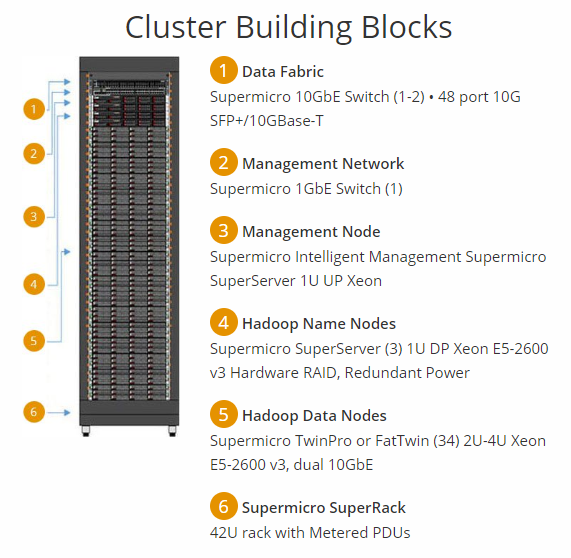


Figure 6 Cluster specification

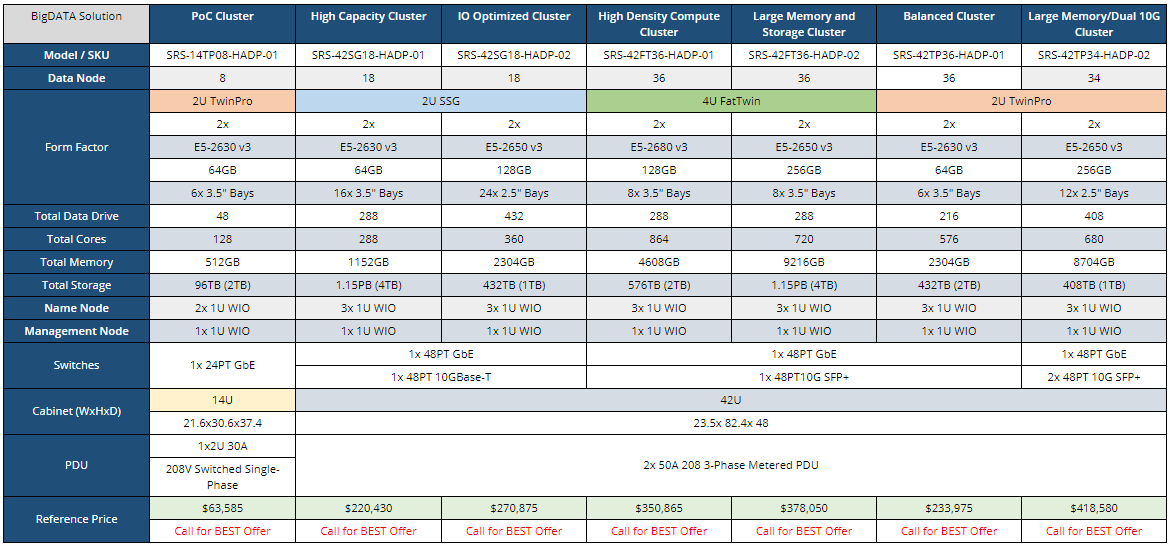


Figure 7 Hardware specification and estimated price

# Timeline

For this project, we proposed a timeline of 4 month until deployment.



# Proof of concept

To validate our framework concept, we installed a single node cluster of the framework into our test machine. The details of the installation as below

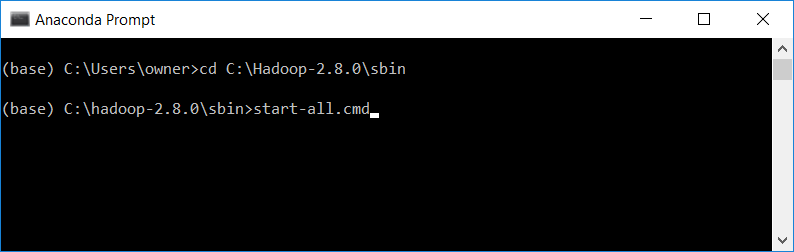
1. Java SDK JDK-8u191
2. Apache Hadoop-2.8.0
3. Apache Spark 2.4.0
4. Anaconda 3 (data analytics source control)
5. Python 3.6 (programming language for data processing)
6. Pyspark (python module for Spark)
7. Jupyter notebook (as python code editor)

For data analysis purposes, Green Taxi NY has provided us with the Green Taxi NY ride data for 2015. This dataset includes trip records from all trips completed in green taxis in NYC in 2015. Records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Livery Passenger Enhancement Program (LPEP). This data set contains 19.2M rows and 21 columns with the data total size of 3.2GB

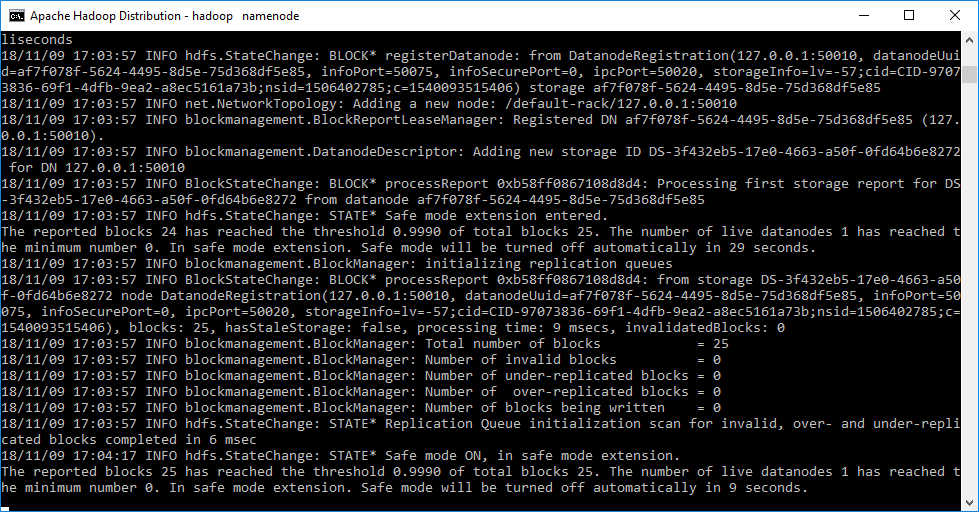
The data set from [here](https://data.cityofnewyork.us/Transportation/2015-Green-Taxi-Trip-Data/gi8d-wdg5)

## Run Hadoop on test machine

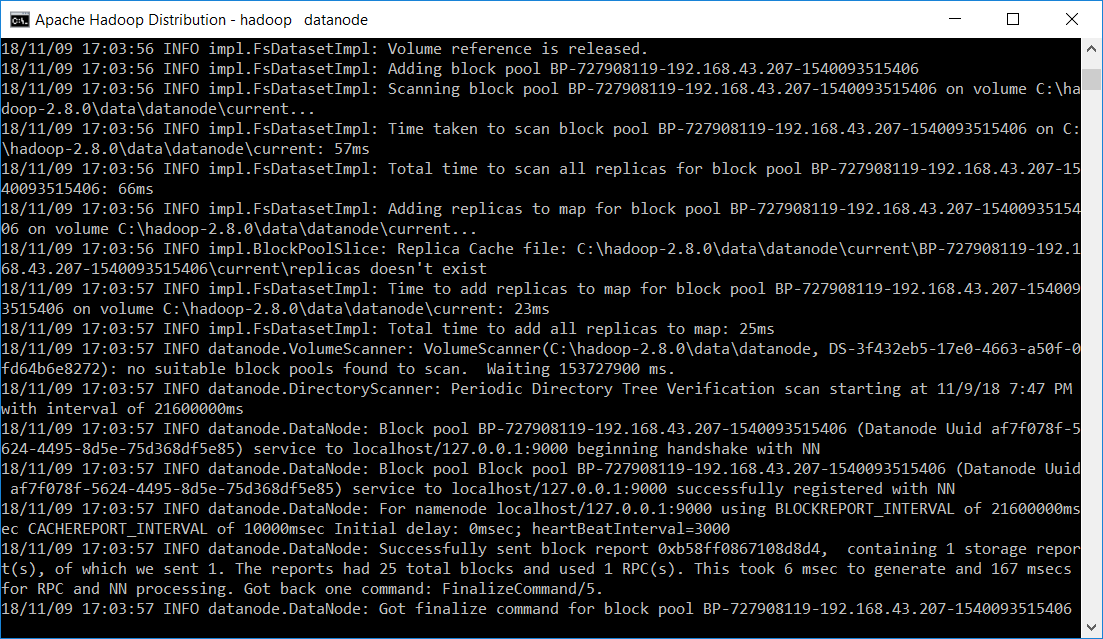
1. Run command to start Hadoop server



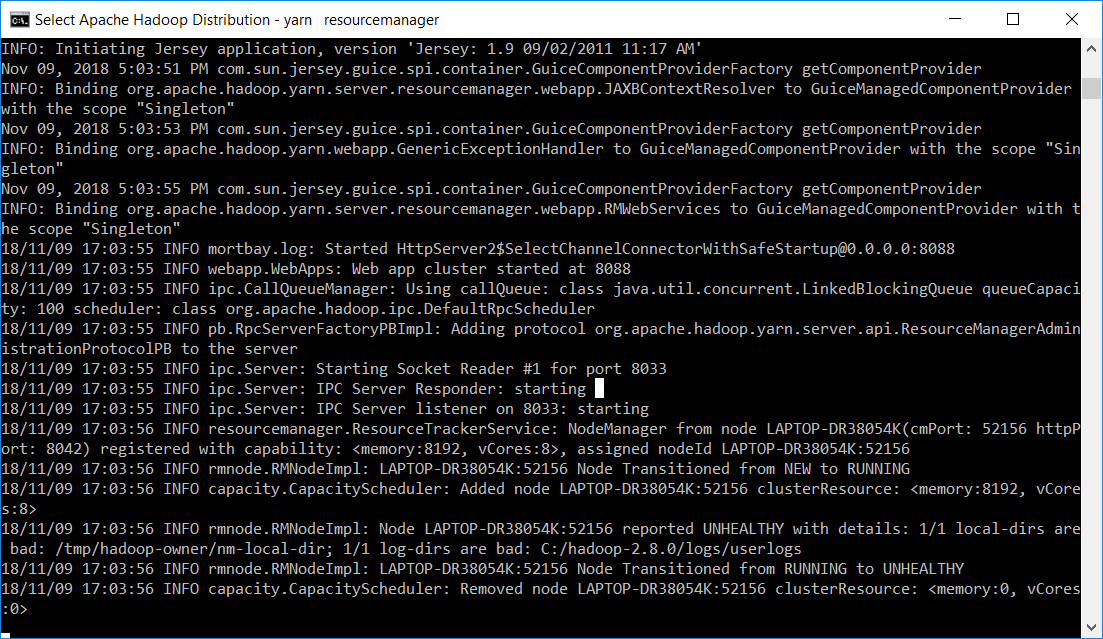
1. Hadoop server running
2. Namenode



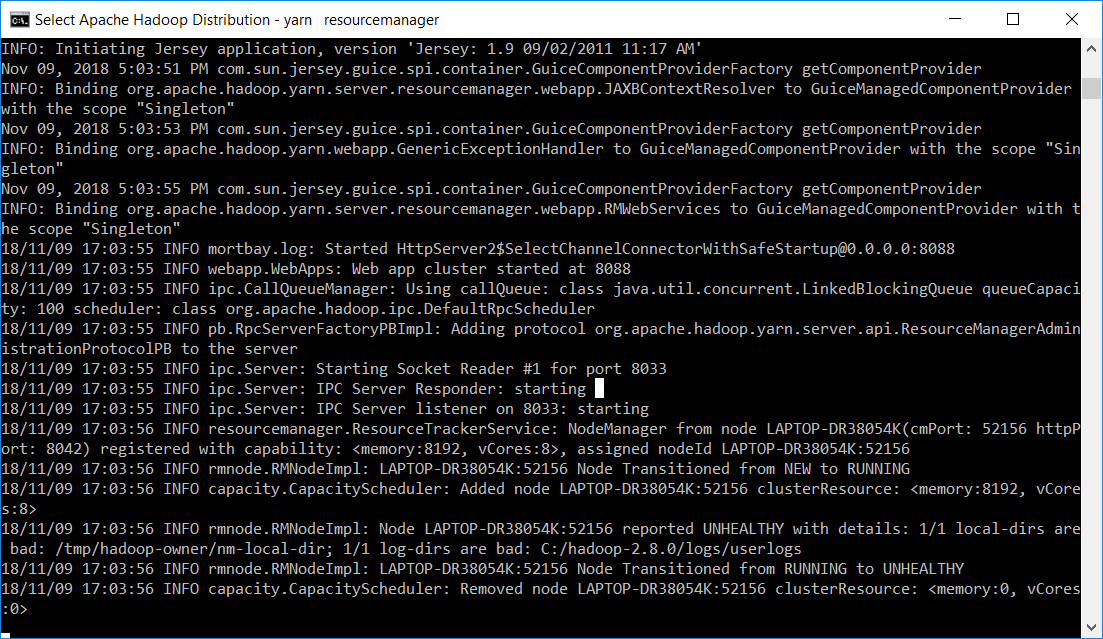
1. Datanode



1. Yarn resource manager

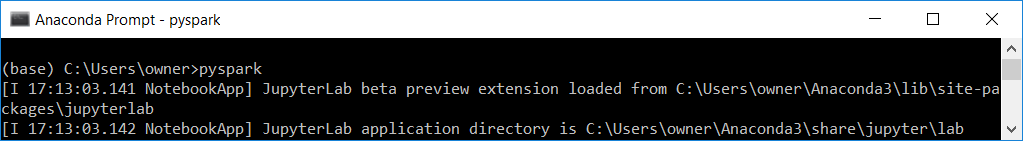


1. Node manager

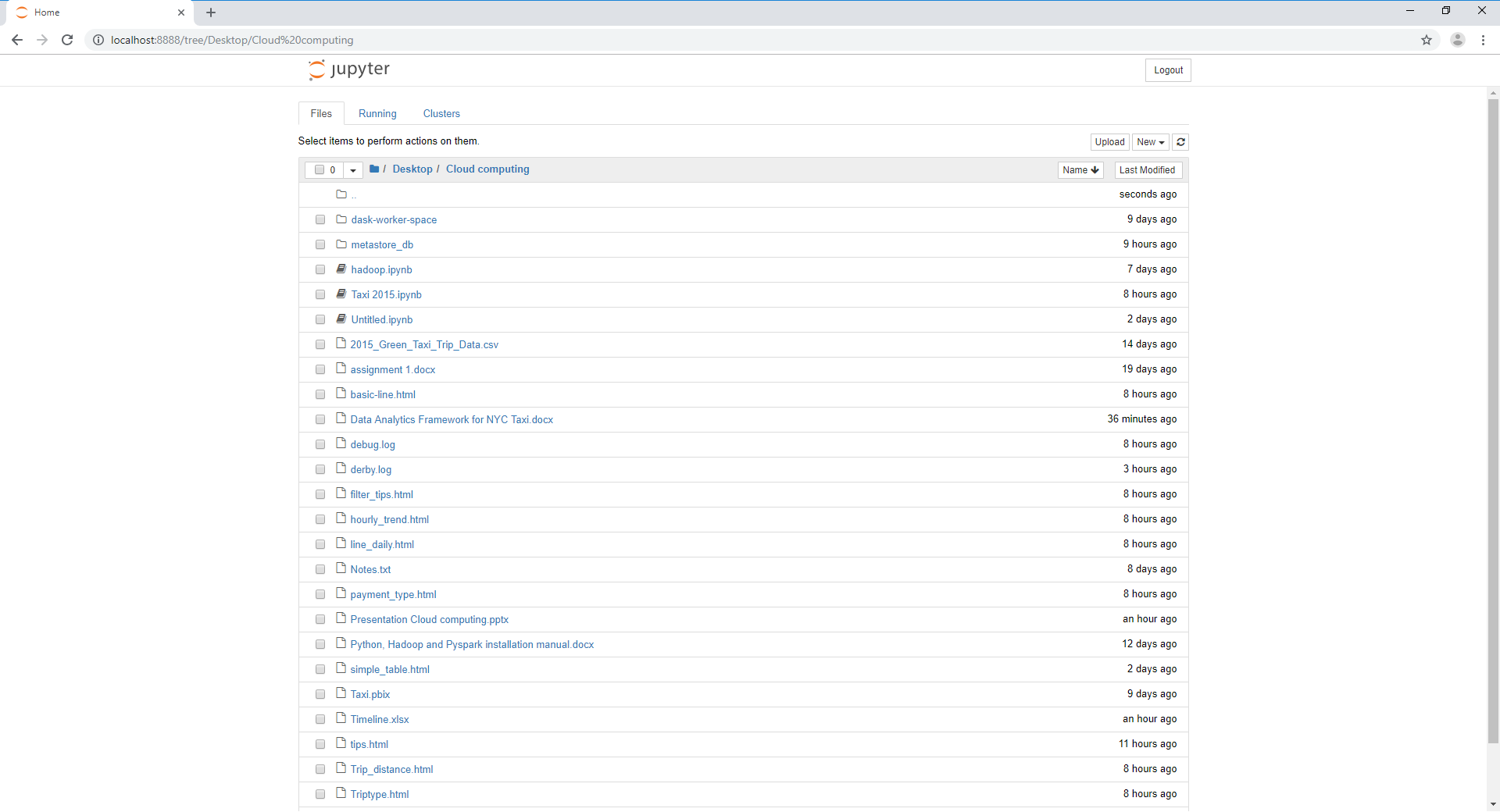


## Run Spark and Jupyter Notebook on test machine

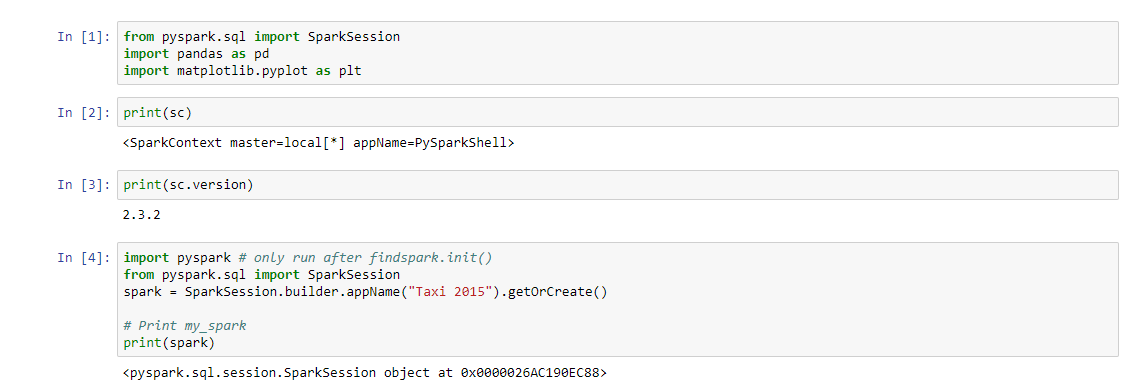
1. Run command to start Apache Spark server



1. Start Jupyter Notebook

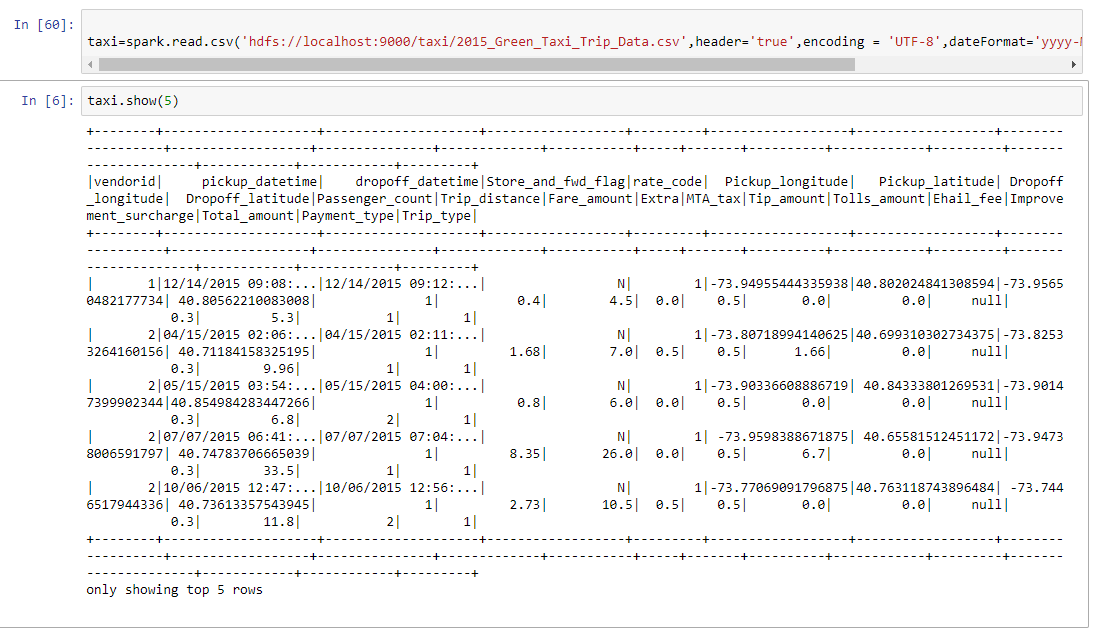


1. Test spark connection from Jupyter Notebook

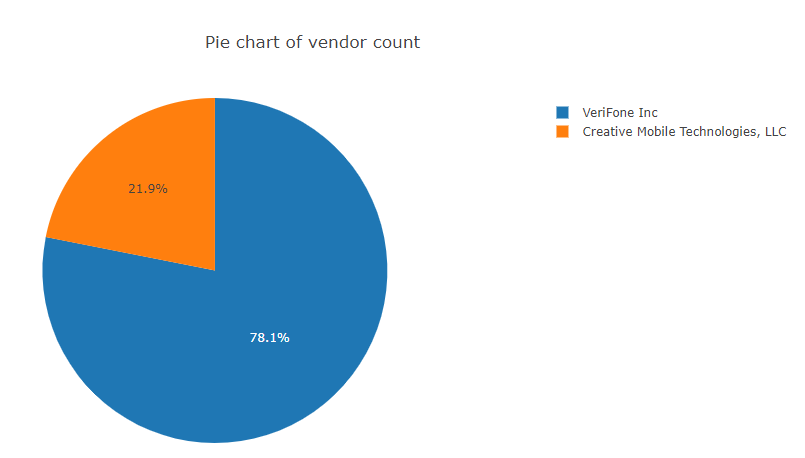


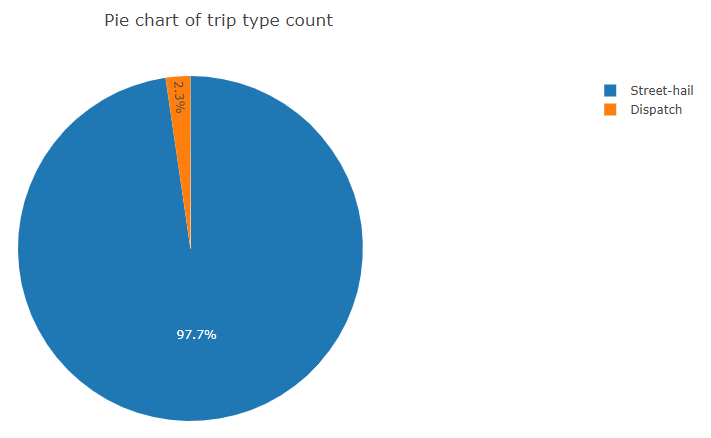
## Use Spark to load NYN Green Taxi 2015 data from Hadoop

1. Load data from Hadoop hdfs file system into Spark dataframe.

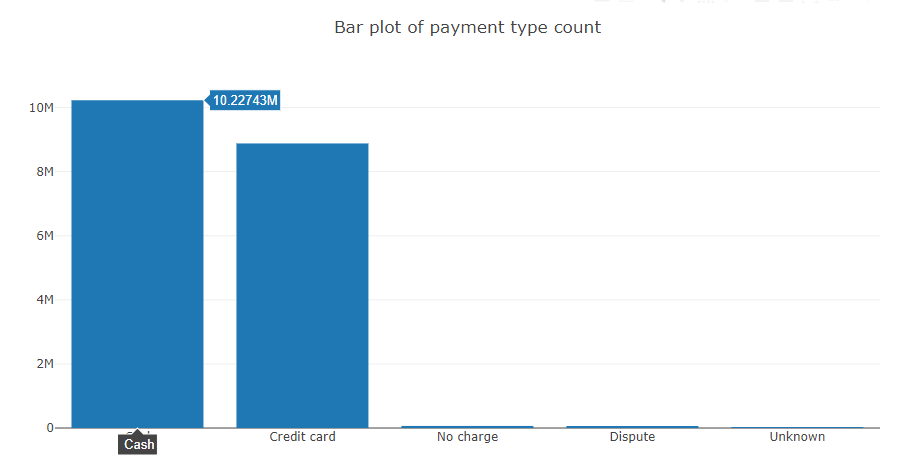


1. Perform data analysis and chart plotting using python program. Analysis sample result as below.
2. Pie chart example

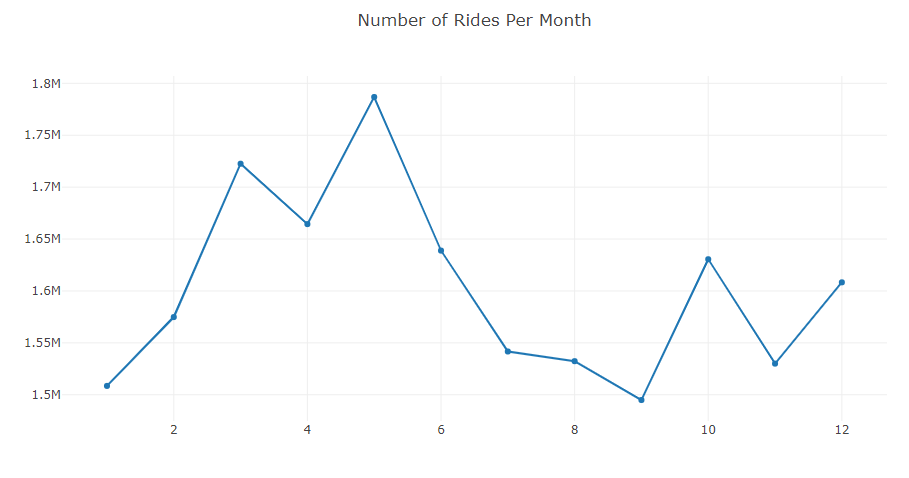


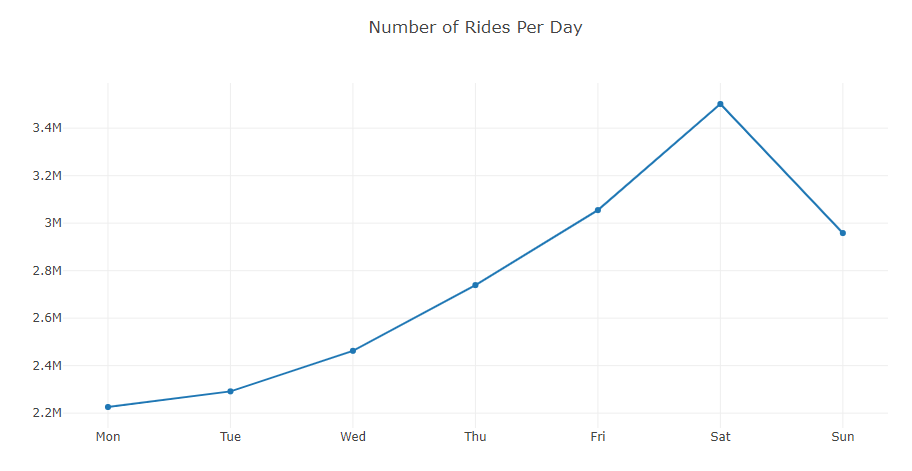


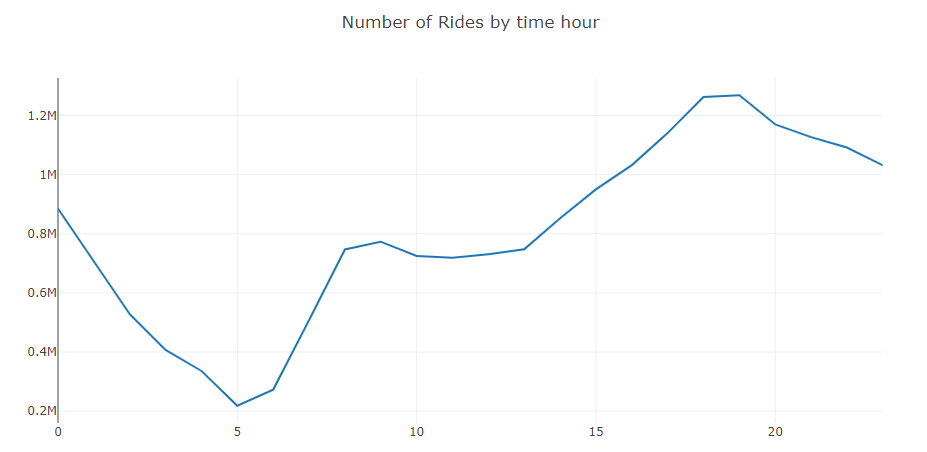
1. Bar chart example



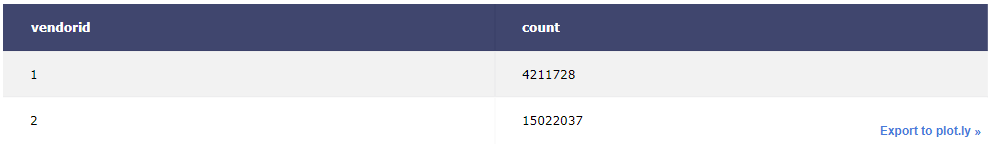
1. Line chart example

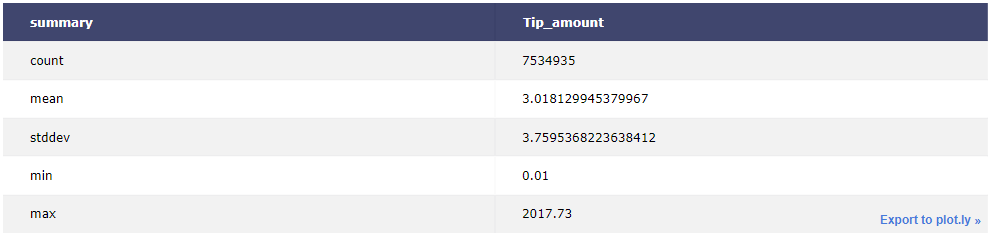


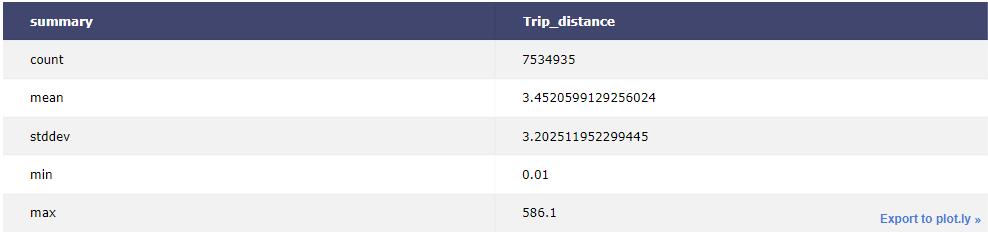




1. Table example







# Costing

Costing for this project has been estimate as in below table.

Table Project estimated cost

|  |  |  |
| --- | --- | --- |
| Cost Group | Details | Estimated cost in $ |
| Hardware/Infrastructures costs | Servers | 63585 |
| Peripherals |
| Network |
| Storage |
| Communication costs | Local Area network | TBD |
| Wide area network | TBD |
| Remote access | TBD |
| Sofware costs | License | TBD |
| Subscription fees | TBD |
| Implementation costs | Development | 100000 |
| Customization |
| Integration |
| Training |
| Consulting |
| Non functional testing |
| Management cost | Hardware upgrades & Software upgrades | TBD |
| Hardware upgrades & Software administration | TBD |
| Legal cost | 1000 |
| Support cost | Support staff | TBD |
| Staff training | TBD |
| Travel | 10000 |
| Support contracts | 1000 |
| Overhead labors | 131833 |
| High Availability cost | TBD |
| Disaster recovery cost | TBD |
| Ticketing and trouble shooting cost | TBD |
| Monitoring cost | TBD |
| Internal audit cost | TBD |
|  | Total | 307418 |

Table Overhead labor calculation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Yearly [$] | Monthly [$] | FTE | Project Estimate |
| Hadoop developer | 90500 | 7542 | 2 | 60333 |
| Data engineer | 90500 | 7542 | 1 | 30167 |
| IT project manager | 100000 | 8333 | 1 | 33333 |
| Support staff | 24000 | 2000 | 1 | 8000 |
|  |  |  | Total | 131833 |

Total estimated cost for this project is $307418.

# Summary

In this report, we have explained the proposed framework that we are going to use for the project. We have also explained on the pro and con of the technology that we are going to use. A proof of concept model has been installed in our test machine and we successfully demonstrated on how Hadoop and Spark can be integrated. We also demonstrated on how data from Hadoop HDFS file can be connected to the analytics tool.

In term of costing, we found out the cost is very competitive and can be done below the budget allocated by NYC Green Taxi which is $400000. Therefore, in our opinion, the project is very viable and we requested a budget of $150000 to kick off the project upon agreement with NYC Green Taxi.