Traffic Change in Portland

from 2017 to 2020

Qiushi Teng

Table of Contents

- **Systems and Techniques**

  Apache Spark and PySpark

- **Data**

  PORTAL, US26 WB, from 2017 to 2020

- **Data Analysis**

  Yearly, Monthly, Different days of a week

PART I

Systems and Techniques

Apache Spark and Pyspark

Apache Spark

A unified analytic engine for large scale data processing

Speed

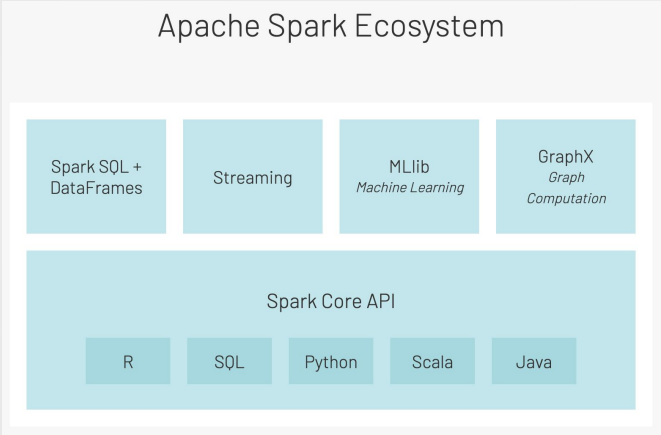
Spark can be 100x faster than Hadoop for large scale data processing by exploiting in memory computing and other optimizations.

Ease of Use

Support APIs in multiple programming languages, including Scala, Java, Python, R and SQL.

Generality / A Unified Engine

Spark comes with libraries including SQL and dataframe, Spark Streaming, MLlib for machine learning, and Graph X, which can be combined in the same application.



PySpark

A Python API for Spark

PySpark features a few libraries, such as:

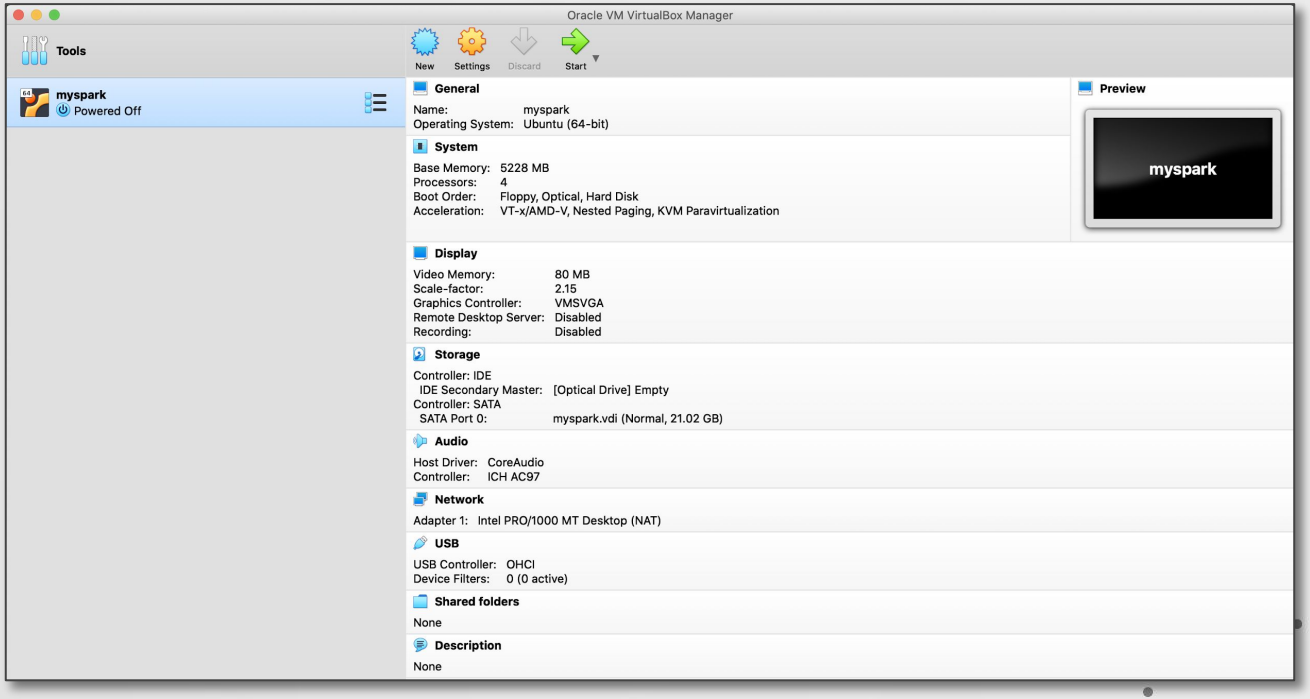
- **PySparkSQL**
  a PySpark library to apply SQL and introduce dataframe
- **MLlib**
  Supporting many machine learning libraries for classification, regression, and more.
- **GraphFrames**
  Graph processing library that provides a set of APIs for performing graph analysis efficiently

Benefits

- Python is easy to learn and implement
- Better readability of code and maintenance
- It features various options for data visualization, which is difficult using Scala or Java

My System Setup

VirtualBox + Ubuntu + Jupyter Notebook



SparkSQL

Apache Spark's module to work with structured data

SparkSQL provides a programming abstraction called DataFrames and can also act as a distributed SQL query engine.

```
In [1]: from pyspark.sql import SparkSession

In [2]: spark = SparkSession.builder.appName('traffic').getOrCreate()

In [3]: df = spark.read.csv('2017-06.csv', inferSchema=True, header=True)

In [4]: df.show()

+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+-----+------+
|          starttime|resolution|detector_id|speed|volume|occupan
cy|countreadings|delay|traveltime|  vht|   vmt|
+-------------------+----------+-----------+-----+------+-------
--+-------------+-----+----------+-----+------+
|2017-06-01 00:00:00|  01:00:00|     100643|50.92|    49|     0.
65|          180|  0.1|      0.68| 0.56| 28.42|
|2017-06-01 00:00:00|  01:00:00|     100874|62.76|   226|     1.
19|          162|-0.05|       1.1| 4.16|261.03|
```

PySparkSQL

Handling Missing Data

There are three ways to handle missing data

1. Drop the missing data
2. Fill in the missing data with 0
3. Fill in the missing data with mean value

```
In [9]: from pyspark.sql.functions import mean

In [10]: mean_val = df.select(mean(df['speed'])).collect()

In [11]: mean_val
Out[11]: [Row(avg(speed)=57.085596295234986)]

In [12]: mean_speed = mean_val[0][0]

In [13]: mean_speed
Out[13]: 57.085596295234986

In [15]: new_table = df.na.fill(mean_speed, ['speed'])

In [20]: new_table.agg({'speed': 'avg'}).show()
+------------------+
|        avg(speed)|
+------------------+
|57.085596295235014|
+------------------+
```

Part II

Data

PORTAL

PORTAL

Official transportation data archive for the Portland-Vancouver Metropolitan region

Dataset Used

- **Yearly Comparison**

  Data of June from 2017 to 2020

- **Monthly Comparison**

  Data from all different months in 2019

- **Comparison between different days of a week**

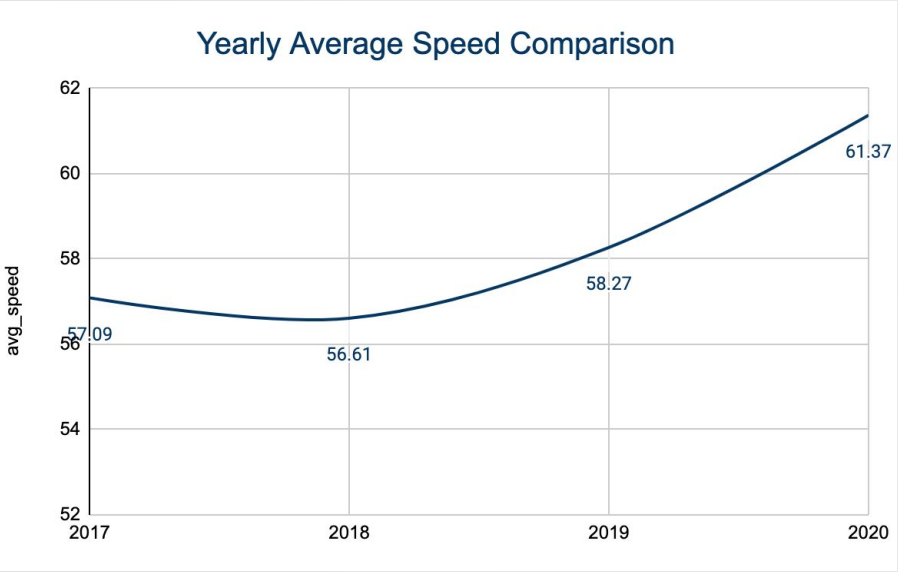
  Data from the first week of June from 2017 to 2020

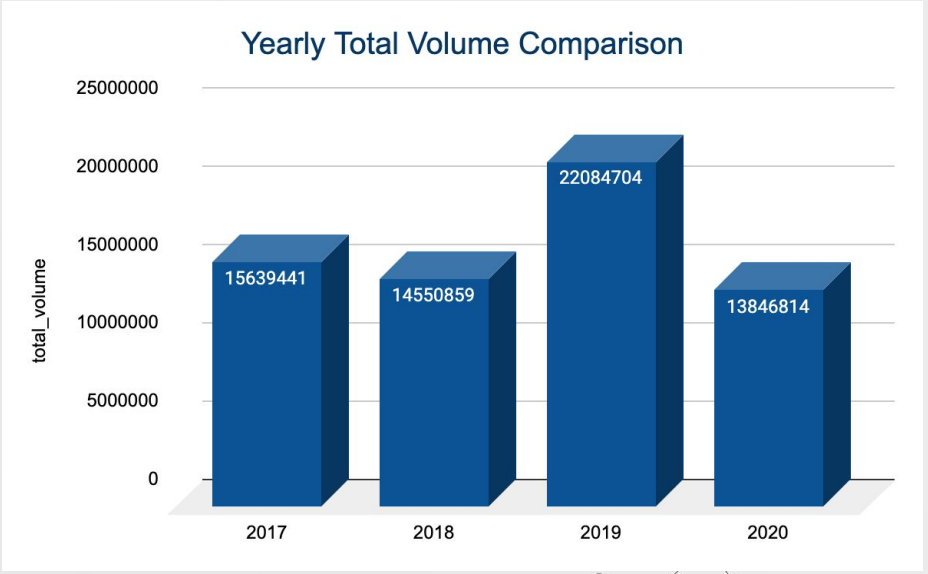
PART III

Data Analysis

Yearly Comparison

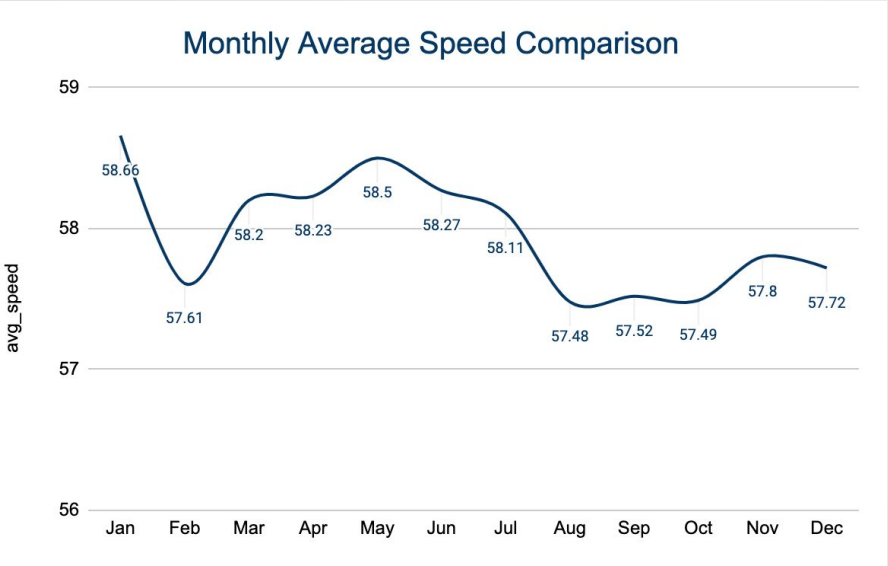
from 2017 to 2020

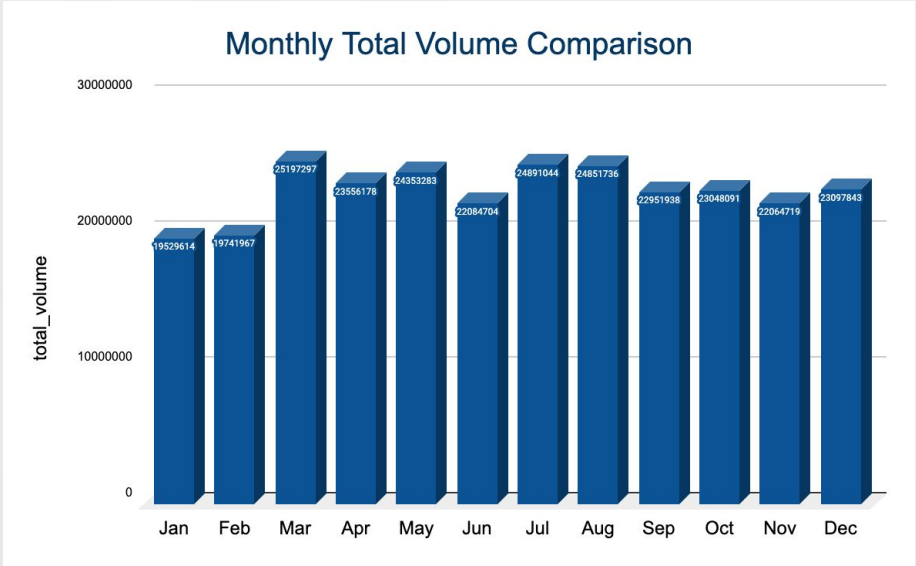




Monthly Comparison

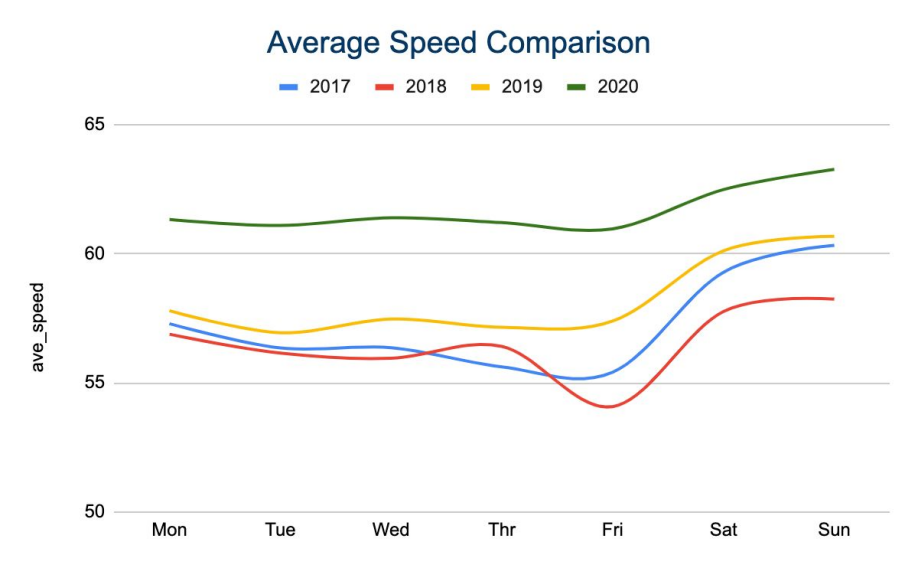
of Year 2019

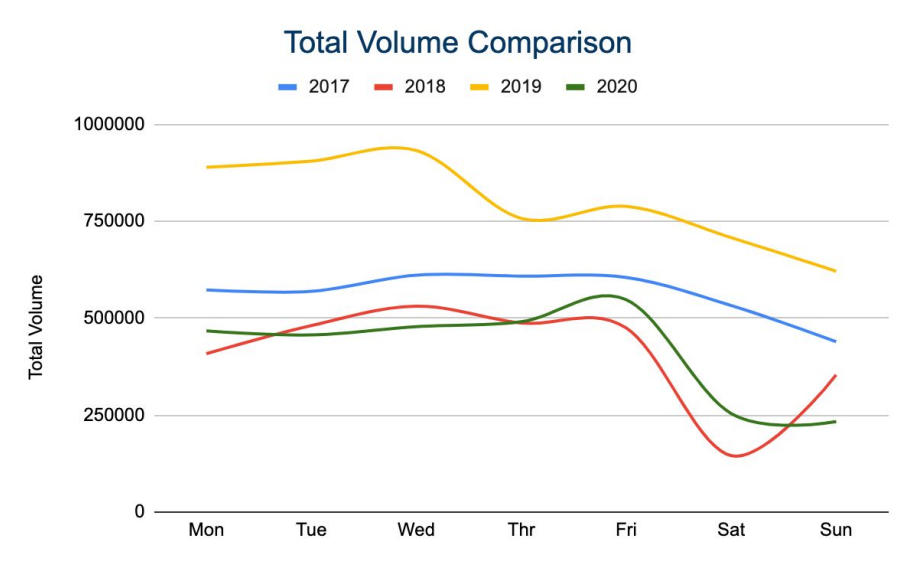




Comparison between days of a week

The first week of June from 2017 to 2020





Thanks

References

Apache Spark

Apache Spark Under The Hood
https://tanthiamhuat.files.wordpress.com/2019/01/apache-spark-under-the-hood.pdf

Getting Started with Apache Spark from Inception to Production
https://mapr.com/ebook/getting-started-with-apache-spark-v2/assets/Spark2018eBook.pdf

Large-scale text processing pipeline with Apache Spark
https://arxiv.org/pdf/1912.00547.pdf

Dataset

From PORTAL
Analysis 1: Jun 2017, Jun 2018, Jun 2019, Jun 2020
Analysis 2: from Jan to Dec in 2019
Analysis 3: 06/05/2017 - 06/11/2017, 06/04/2018 - 06/10/2018, 06/03/2019 - 06/09/2019, 06/01/2020 - 06/07/2020