

EURO

HPDA Experience in Public Transportation Domain
Abdulkadir Kılavuz, Mehmet Barutcu, Cevat Şener (METU)

Outline

parabol



- Introduction
- Case Study - 1 Origin-Destination
 - Problem Definition
 - Solutions
- Case Study - 2 Timetable Optimization
 - Problem Definition
 - Solution

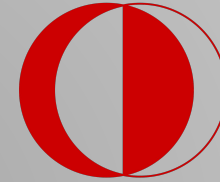
- Results & Benefits

Collaborators

Introduction



- TÜBİTAK ULAKBİM
- Middle East Technical University
 - Dr. Cevat Şener
- Parabol Yazılım
 - Abdulkadir Kılavuz, Computer Eng.
 - Mehmet Barutcu, Industrial Eng.



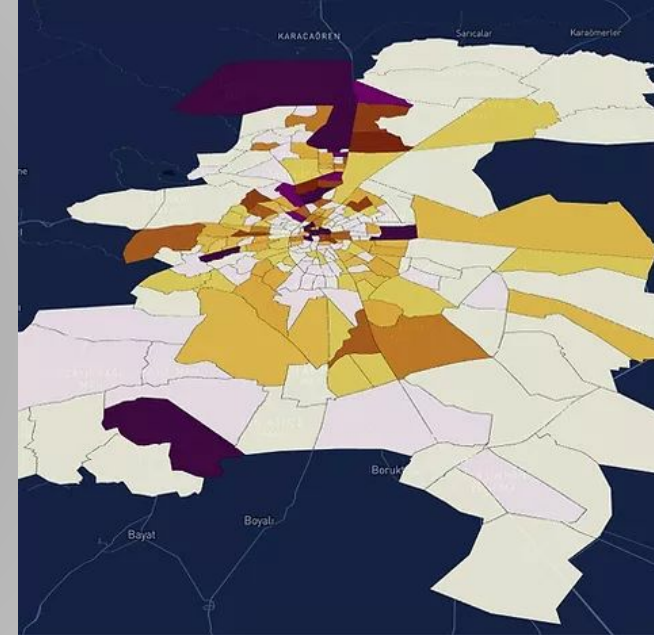
Parabol Yazılım

Introduction

parabol



- Parabol Yazılım (paraboly.com) has been carrying out R&D activities in the field of intelligent transportation systems since 2011.
- The company is also developing a public transport analysis platform called Cermoni (cermoni.app), which uses the outputs of these case studies.
- Passenger card and bus location data were used.
- In Konya Metropolitan Municipality, spatiotemporal data of ~10M rows per month is processed.

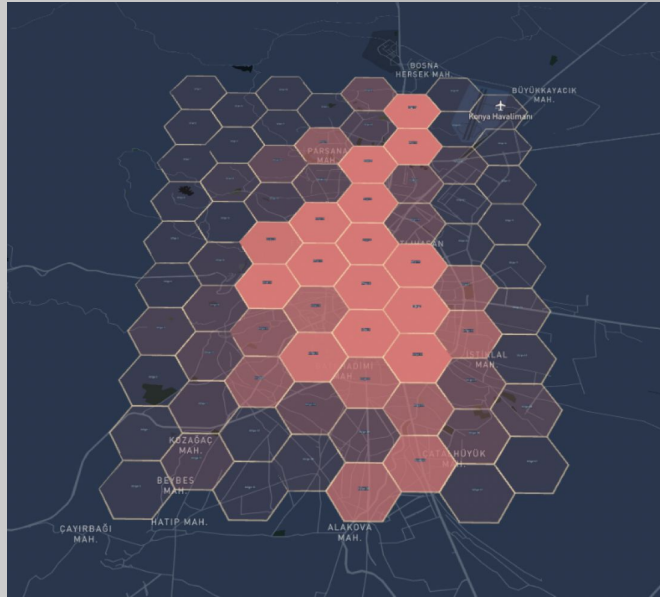


Problem Definition for Case Study-1

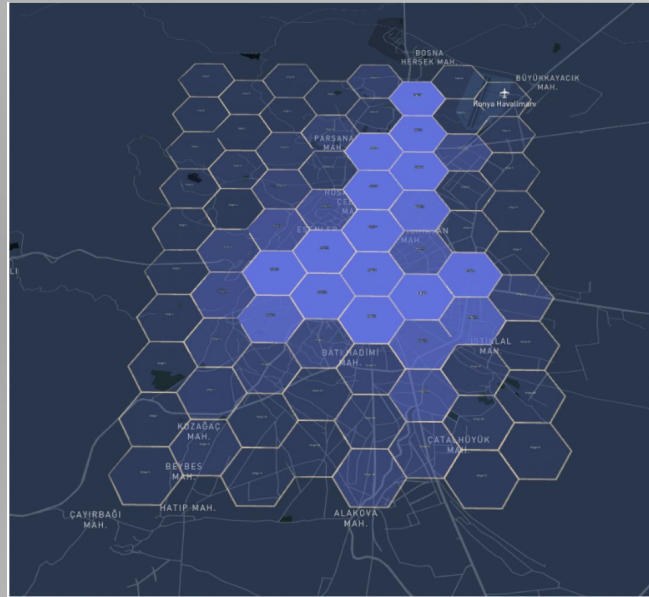
- The public transport analysis platform Cermoni needs to analyze the patterns of passengers.
- It is desired to determine the demand for public transportation in the city.
- In order to determine the demand, it is necessary to determine the regular journeys and multidimensional (passenger, time, space) clustering.
- With the analyzes made, the starting and ending points of the journeys were analyzed according to the regions (OD analysis).



Problem Definition



Origin points at Morning



Destination points at Morning

Problem Definition

Need for HPC



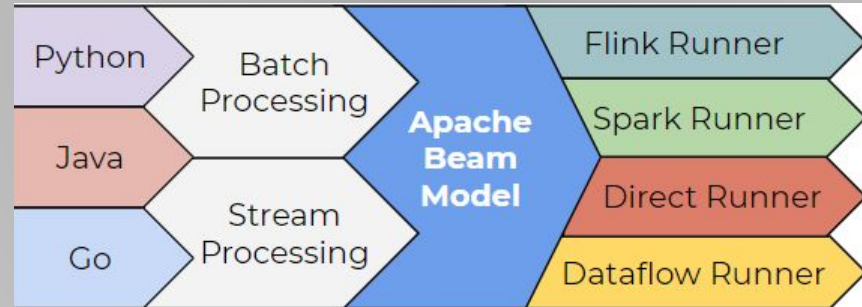
- Since clustering-based analyzes in space (coordinate system), time and passenger dimensions will be performed on a data over 10M rows, the resource requirement is high.
- Parallel operation is required for spatiotemporal analyses.
- Parabol normally develops algorithms on the company's own servers or in environments created on GCP.
- The suitability, advantages and disadvantages of the TRUBA environment for such a big data problem are examined.

First Solution Approach

Apache Beam



- Unified programming model - *“Write once, run anywhere”*
- Stack and Stream data processing
- Runner: Google Dataflow, Flink, Spark, Samza etc.
- Java, Python & Go



First Solution Approach

Apache Beam ... and Problems



- Developing on Apache Beam by utilizing Parabol Software's previous knowledge
- Running the Beam model to be obtained on the Spark cluster



- Its promises were good, but with Python, it failed to work on Spark.
- Open source support is very poor.
- Running outside of Google Cloud Dataflow is practically problematic.

Second Solution

Apache Spark Experience



- It was decided to use Apache Spark.
- We have adapted the algorithms we use for Apache Beam to work on Spark with minor changes.
- Also, Spark needed to be ready to run on TRUBA.
- After it was installed on TRUBA, we experimented with Spark's RDD, Dataframe and SparkSQL APIs.



Steps



- We reviewed the documentation on TRUBA and Slurm and did some minor tests.
- With the support of the TRUBA Technical Team (especially Dr. Kerem Kayabay), Spark was made available on TRUBA; attempts were made; problems were fixed.
- We have adapted the algorithms we use for Apache Beam to work on Spark with minor changes.



TRUBA with Spark

Apache Spark Experience



- Anaconda is supported.
- Spark clusters are submitted as Slurm job.
- Then the job to be run is sent to the cluster with spark-submit.
- Adding Spark jars is very simple.
- Setting the necessary parameters for Slurm and Spark is a bit of a challenge at first.
- The waiting time in the queue can be long.
 - A Trial and Error development cycle is not very easy to implement.



Problem Definition for Case Study-2



- The second stage of the public transportation decision support mechanism is to determine the optimal number of frequencies for different trips related to the same terminals.
- The timetable optimization algorithm in the Cermoni platform (*cermoni.app*) provides the optimal frequencies for public transport agencies.
- Algorithm tries to find the solution that best meets the OD requirement of the city with the number of vehicles and drivers available.
- Optimization system uses origin destination information coming from the public transport analysis platform, as well as the route and terminal information and user specific criterias.

Problem Definition for DASK

- With the analyzes made, the optimal frequencies are determined.
- Requires all terminal based routes in order to get optimal results.

HOCAFAKIH TIP FAKÜLTESİ ANASULTAN	MERAM YAKA ANASULTAN	MERAM YENİYOL DÖRTOKKA
06:00	06:00	06:00
06:05	06:06	06:03
06:10	06:12	06:06
06:15	06:18	06:09
06:20	06:24	06:12
06:25	06:30	06:15
06:30	06:36	06:18
06:35	06:42	06:21
06:40	06:48	06:24
06:45	06:54	06:27
06:50	07:00	06:30
06:55	07:06	06:33
07:00	07:12	06:36
07:05	07:18	06:39
07:10	07:24	06:42
07:15	07:30	06:45
07:20	07:36	06:48
07:25	07:42	06:51
07:30	07:48	06:54
07:35	07:54	06:57

Steps



- We went through the documentation on TRUBA and did some short tests.
- With the support of TRUBA Technical Team, special thanks to Dr. Kerem Kayabay, the singularity image was created with required python libraries and uploaded to TRUBA.
- We ported the algorithm onto Dask with minor modifications.
- The study have not finished yet. We still test our algorithm in TRUBA with challenging parameters.



TRUBA with DASK

Dask Experience



- Singularity is supported, which enables portable and repeatable execution of complex applications on HPC clusters.
- It is easy to build singularity images with required conda packages.
- Scripts are submitted as Slurm job by using singularity images.
- Waiting time in the queue is short.



Results & Benefits



- With the collaboration of Parabol + TRUBA Technical Team + METU, the TRUBA infrastructure has proven to be a suitable platform for big data algorithms.
- HPDA enables us to run our solutions in parallel and get the results in minutes rather than hours.
- Outputs that will contribute to public transport service quality are obtained in a meaningful period of time.

Thank you...



This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 951732. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, United Kingdom, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Switzerland, Turkey, Republic of North Macedonia, Iceland, Montenegro