“Implementation of K-Anonymity for the Use Case of Automotive Industry in Big Data Context.”

Sharada Sowmya, 08.03.2021 – Master Thesis Final Presentation

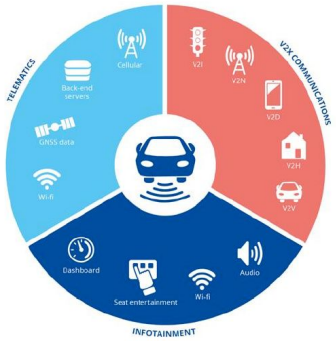
Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
wwwmatthes.in.tum.de

Outline

- Motivation and Problem Statement
- Use Case
- Foundation
- Research Questions
- Outcome of Literature Review
- Data Preparation
- Implementation
- Results and Evaluation
- Conclusion and Future Work

© sebis

Motivation and Problem Statement



- **Connected Car** is a vehicle that is equipped with internet access enabling it to share data.
- This includes sensitive driver and passenger information including driving behavior, biometric, health, location and communication data.

- Analytics on the behavioral car datasets provides an opportunity for improved user experience as well as to diversify business models.
- However, presence of sensitive information poses a risk to user privacy which can be mitigated through data anonymization.

© sebis

Use Cases

- The collected car data can be usage-based data such as mileage or event based data such as charging status of the car captured depending of whether the car is being charged or not.
- Having consent of the car owner, this data can be made accessible for corporate third parties.
- Two use cases for third party service providers were found post interviews conducted at BMW.

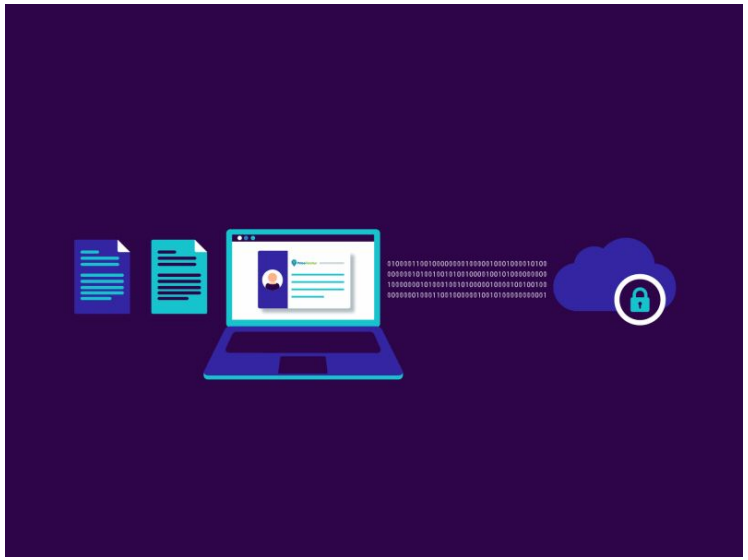
Use case	Focus	Outcome
Advertisement	Smart placement of billboards based on drivers driving patterns like frequent traversed location and more stoppage time.	The most traversed path is extracted from dataset without revealing location or identity of car/driver.
Smart charging stations	Smart placement of charging stations based on drivers driving patterns like the most frequent places where the car is likely to run out of charge.	The optimal location of the smart charging station is derived without revealing the location or identity of the car/driver.

© sebis

Foundations

De-identification: “method for transforming a *dataset* with the objective of reducing the extent to which information is able to be associated with individual *data principles*” (ISO, 2018)

K-anonymity: “A table satisfies k-anonymity if every record in the table is indistinguishable from at least *k-1* other records with respect to every set of quasi-identifier attributes” (Sweeney, 2002)



How is this helpful?

- allows data to be used without the possibility of sensitive data being identified.
- with minimal information loss, analytics can be performed on the dataset to return valid outcomes.

© sebis

Research Questions

RQ 1 What are the properties of current k-anonymity implementations/algorithms?

Extensive Literature Research

Analysis of the implementations/algorithms

→ Guideline for usage of k-anonymity algorithms/implementation

RQ 2 What are the requirements for k-anonymity implementations in big data context?

Literature Research on Big Data Anonymization

Talks with Rohde & Schwarz and BMW

→ List of Requirements

RQ 3 How can a k-anonymity implementation in the context of big data look like?

Data Preparation

Implementation

Evaluation of Results

→ Proof-of-concept

© sebis

Outcome of Literature Review

Which technique provides the best anonymization? Depends on a use case basis.

- **Mondrian** → better suited for numerical data sets.
- **Incognito** → for data sets with low number of quasi-identifiers.
- **Datafly** → works better with increase in size of data set.

Distributed vs Centralized:

- In the context of large data sets, a distributed approach is preferred over a centralized one.
- '*partition, anonymize and integrate*' approach with anonymization done in asynchronously is one of the way of handling large data sets without running into memory issues.

ARX Library

- **ARX** cross platform anonymization tool for applying k-anonymity was chosen.
- Combination of Mondrian and Incognito algorithms in a top-down specialization manner.
- Scalable for 50 identifiers which suited our use case.

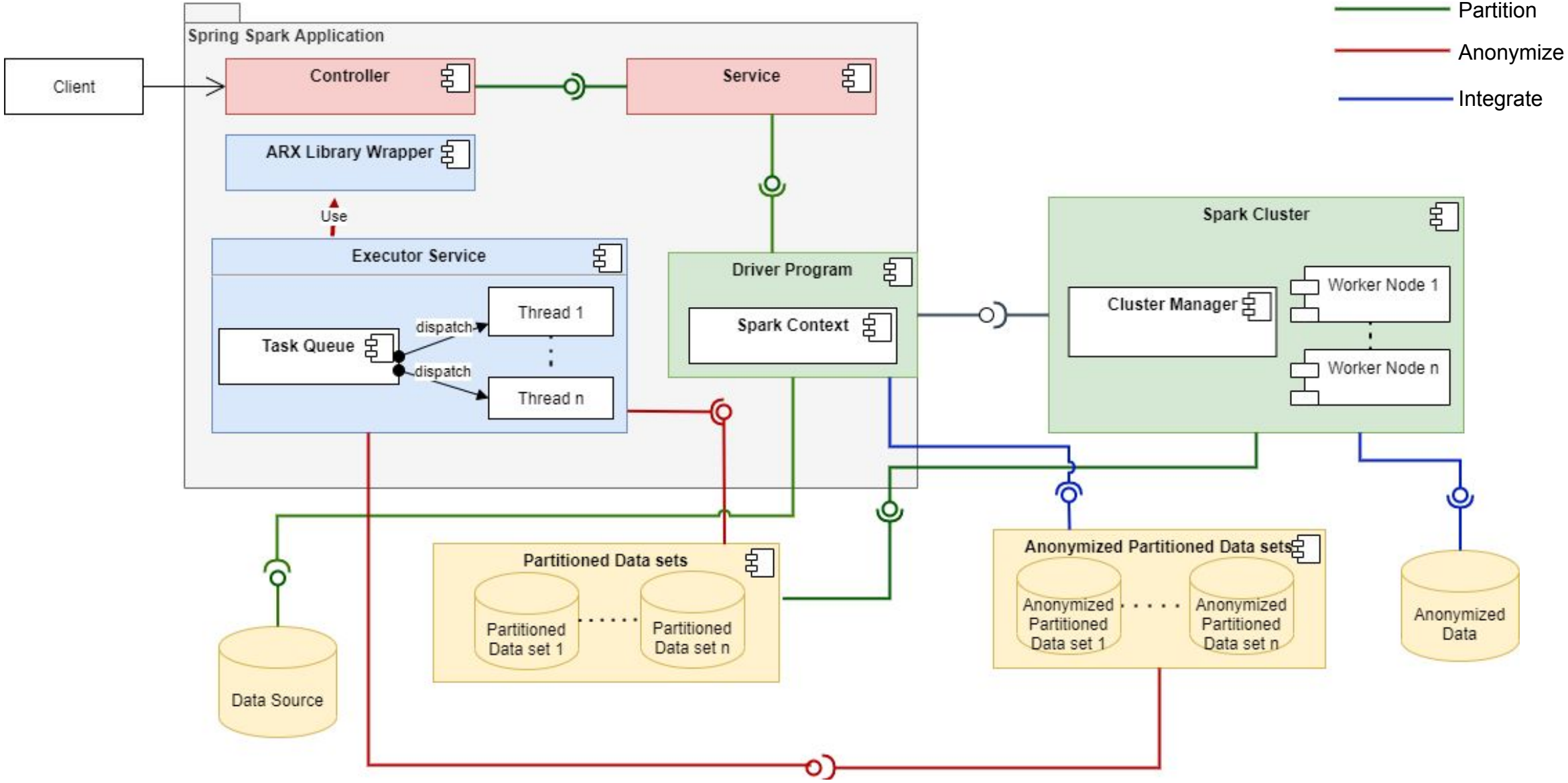
© sebis

Data Preparation

Attribute Name	Attribute value	Identifier type
car_id	85a0e584-12f1-47e0-8a5b-08e19c235a12	Personal Identifier
car_model	5 Series	Insensitive Attribute
charging_method	AC_TYPE1PLUG	Insensitive Attribute
charging_status	CHARGING_ACTIVE	Quasi Identifier
smart_charging_status	RENEWABLE_OPTIMIZED	Insensitive Attribute
fuel_percentage	72.5	Quasi Identifier
mileage	224000	Insensitive Attribute
isc_timestamp	2018-02-16 03:49:56.009	Quasi Identifier
gps_lat	3.2485078566258347	Quasi Identifier
gps_long	5.583555183545527	Quasi Identifier
temperature_external	16	Insensitive Attribute

© sebis

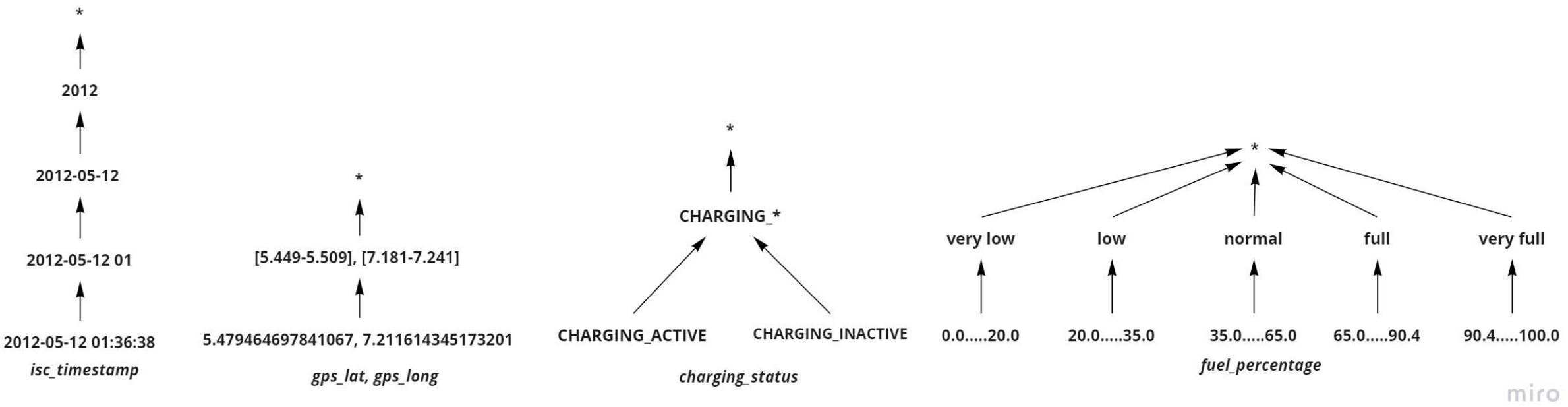
Software architecture diagram



© sebis

Implementation - Anonymization Stratergy

- ARX uses a highly efficient globally-optimal search algorithm for transforming data.
- The transformation of attribute values is implemented through domain generalization.
- Below are the 4 examples of generalization strategy opted for attributes in dataset.



© sebis

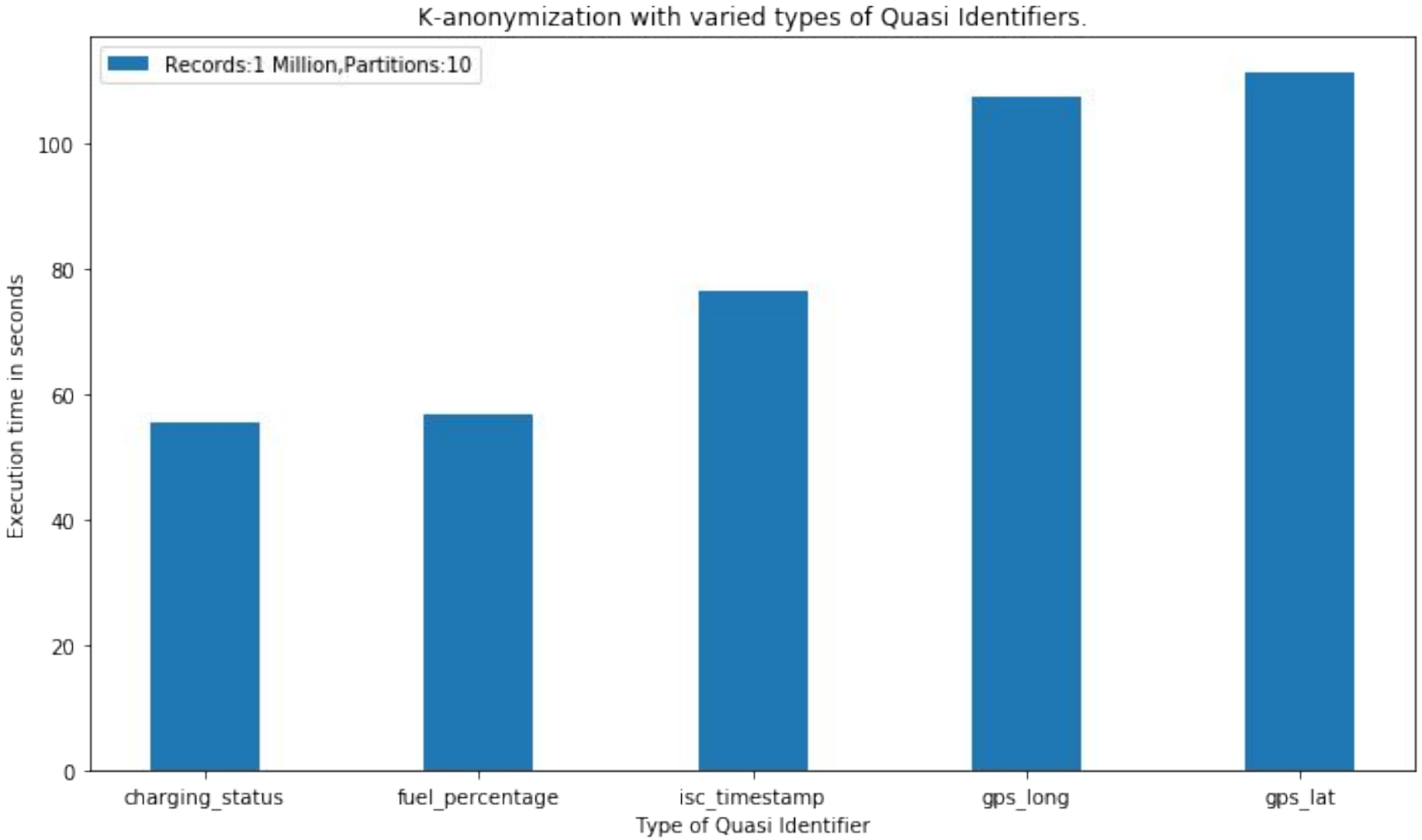
Snapshot of Anonymized Dataset

```
car_id;car_model;charging_method;charging_status;smart_charging_status;mileage;fuel_percentage;isc_timestamp;gps_lat;gps_long;temperature_external;
da516a59-435c-4f1c-9ebd-c8b9f7bca489,1 Series,AC_TYPE1PLUG,CHARGING_INACTIVE,RENEWABLE_UNOPTIMIZED,60000,100.0,2016-09-03 03:05:37,5.479464,7.211614,1
da516a59-435c-4f1c-9ebd-c8b9f7bca489,1 Series,AC_TYPE1PLUG,CHARGING_INACTIVE,RENEWABLE_UNOPTIMIZED,60010,99.0,2016-09-03 03:15:37,5.539464,7.271614,3
da516a59-435c-4f1c-9ebd-c8b9f7bca489,1 Series,AC_TYPE1PLUG,CHARGING_INACTIVE,RENEWABLE_UNOPTIMIZED,60020,98.0,2016-09-03 03:25:37,5.579484,7.331634,5
da516a59-435c-4f1c-9ebd-c8b9f7bca489,1 Series,AC_TYPE1PLUG,CHARGING_INACTIVE,RENEWABLE_UNOPTIMIZED,60030,97.0,2016-09-03 03:35:37,5.639494,7.391644,-3
da516a59-435c-4f1c-9ebd-c8b9f7bca489,1 Series,AC_TYPE1PLUG,CHARGING_INACTIVE,RENEWABLE_UNOPTIMIZED,60040,96.0,2016-09-03 03:45:37,5.699504,7.451654,10
b7b5efad-aede-4e6a-b2cb-4be0e24227fb,5 Series,AC_TYPE2PLUG,CHARGING_INACTIVE,RENEWABLE_UNOPTIMIZED,91050,77.0,2020-02-14 18:05:23,7.372194,8.104344,-2
b7b5efad-aede-4e6a-b2cb-4be0e24227fb,5 Series,AC_TYPE2PLUG,CHARGING_INACTIVE,RENEWABLE_UNOPTIMIZED,91060,76.0,2020-02-14 18:15:23,7.432204,8.164354,-1
b7b5efad-aede-4e6a-b2cb-4be0e24227fb,5 Series,AC_TYPE2PLUG,CHARGING_INACTIVE,RENEWABLE_UNOPTIMIZED,91070,75.0,2020-02-14 18:25:23,7.492214,8.224364,22
b7b5efad-aede-4e6a-b2cb-4be0e24227fb,5 Series,AC_TYPE2PLUG,CHARGING_INACTIVE,RENEWABLE_UNOPTIMIZED,91080,76.0,2020-02-14 18:35:23,7.552224,8.284374,36
```

```
car_id;car_model;charging_method;charging_status;smart_charging_status;mileage;fuel_percentage;isc_timestamp;gps_lat;gps_long;temperature_external;
*,1 Series,AC_TYPE1PLUG,CHARGING_*,RENEWABLE_UNOPTIMIZED,60000,full,2016-09-03,[5.449-5.509],[7.181-7.241],1
*,1 Series,AC_TYPE1PLUG,CHARGING_*,RENEWABLE_UNOPTIMIZED,60010,full,2016-09-03,[5.509-5.579],[7.241-7.301],3
*,1 Series,AC_TYPE1PLUG,CHARGING_*,RENEWABLE_UNOPTIMIZED,60020,full,2016-09-03,[5.549-5.609],[7.301-7.371],5
*,1 Series,AC_TYPE1PLUG,CHARGING_*,RENEWABLE_UNOPTIMIZED,60030,full,2016-09-03,[5.679-5.609],[7.361-7.421],-3
*,1 Series,AC_TYPE1PLUG,CHARGING_*,RENEWABLE_UNOPTIMIZED,60040,full,2016-09-03,[5.669-5.729],[7.411-7-481],10
*,5 Series,AC_TYPE2PLUG,CHARGING_*,RENEWABLE_UNOPTIMIZED,91050,normal,2020-02-14,[7.342-7.401],[8.074-8.134],-2
*,5 Series,AC_TYPE2PLUG,CHARGING_*,RENEWABLE_UNOPTIMIZED,91060,normal,2020-02-14,[7.402-7.462],[8.134-8.194],-1
*,5 Series,AC_TYPE2PLUG,CHARGING_*,RENEWABLE_UNOPTIMIZED,91070,normal,2020-02-14,[7.469-7.529],[8.194-8.259],22
*,5 Series,AC_TYPE2PLUG,CHARGING_*,RENEWABLE_UNOPTIMIZED,91080,normal,2020-02-14,[7.522-7.582],[8.154-8.654],36
```

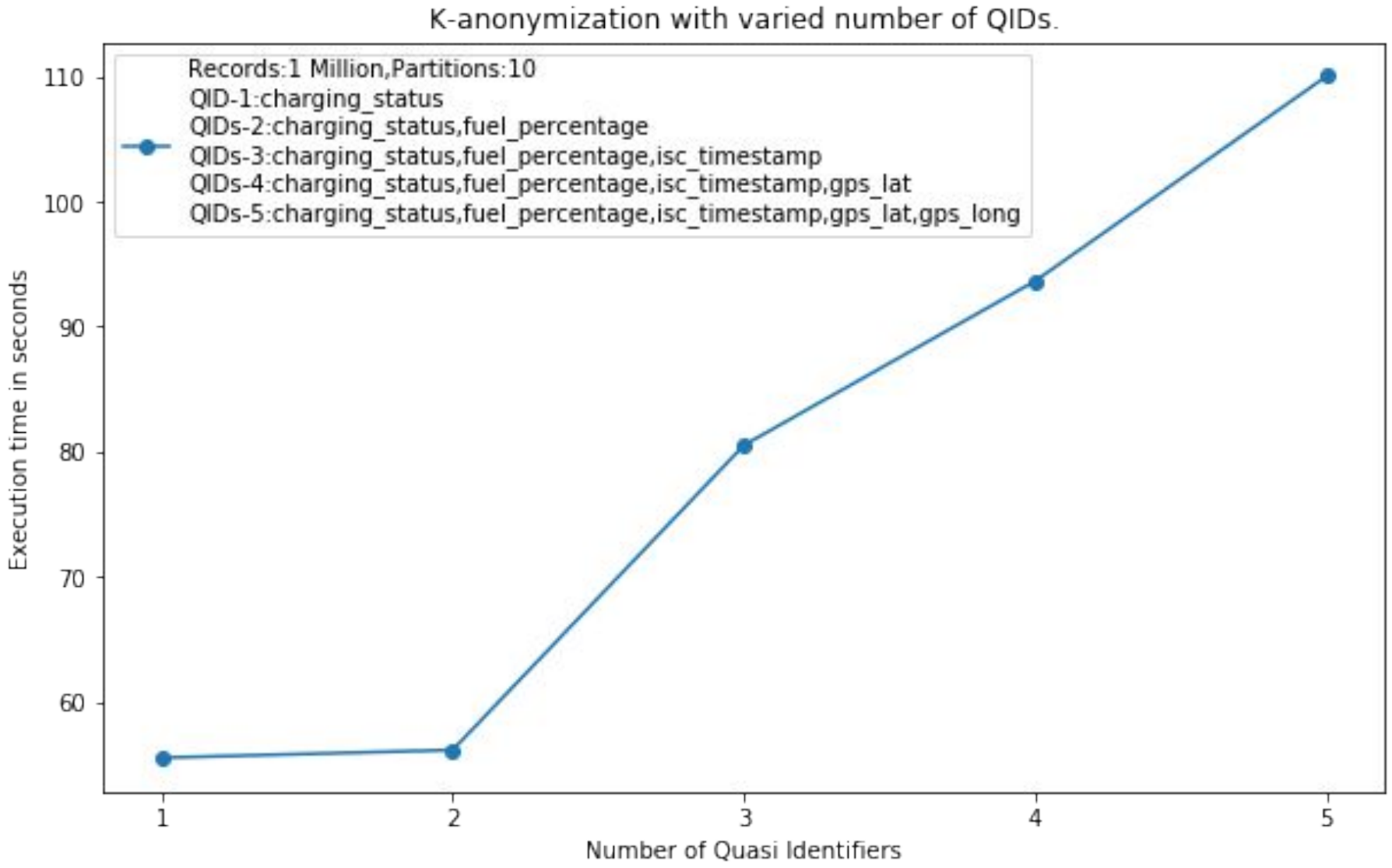

© sebis

Results(1/7) : Execution time vs Quasi Identifier Type.



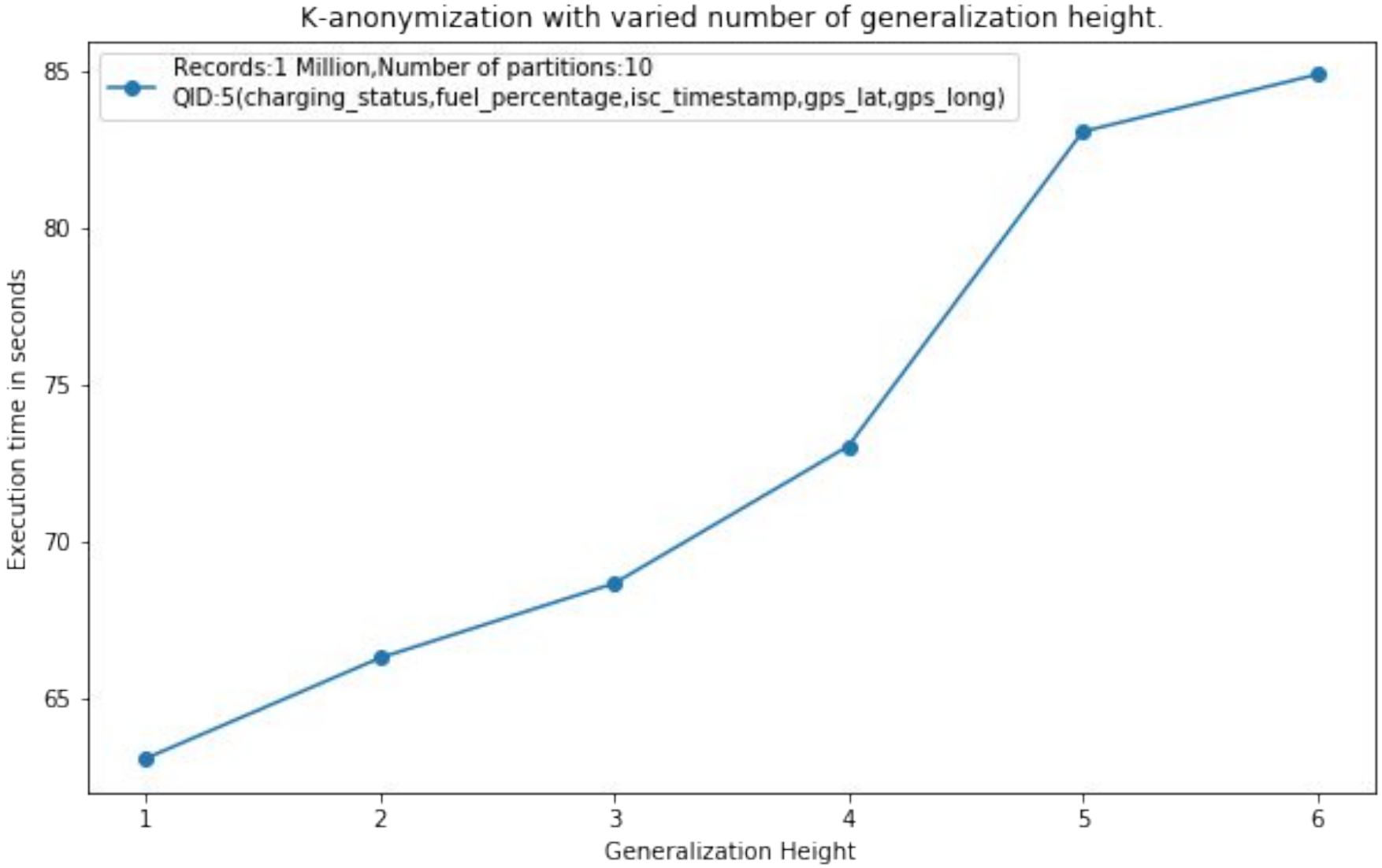
© sebis

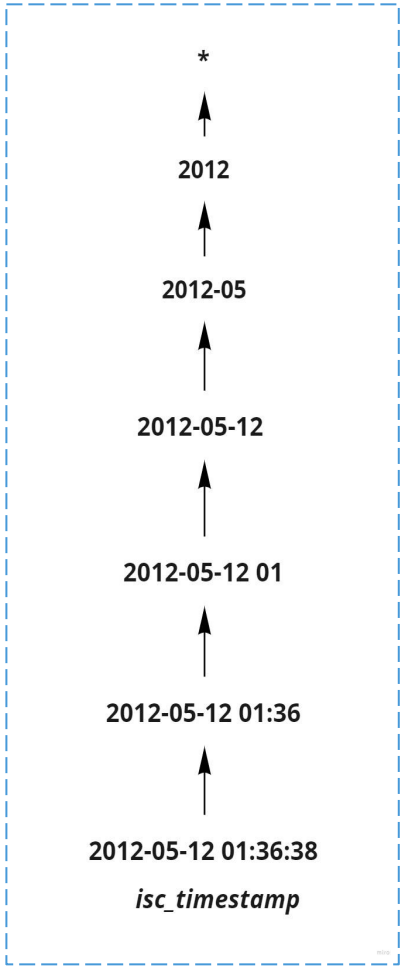
Results(2/7) : Execution time vs Number of Quasi Identifiers.



© sebis

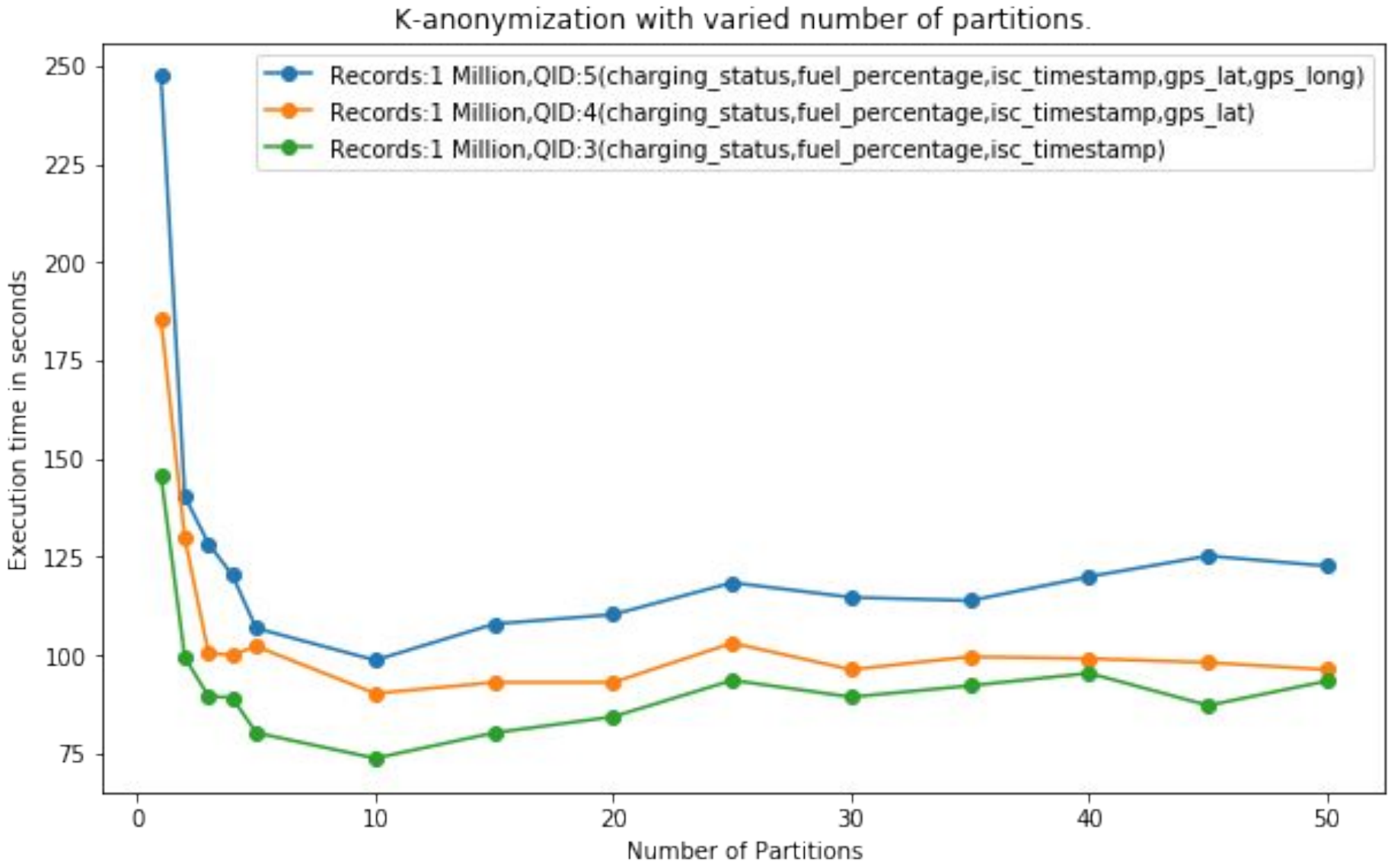
Results(3/7) : Execution time vs Generalization Height of QID.





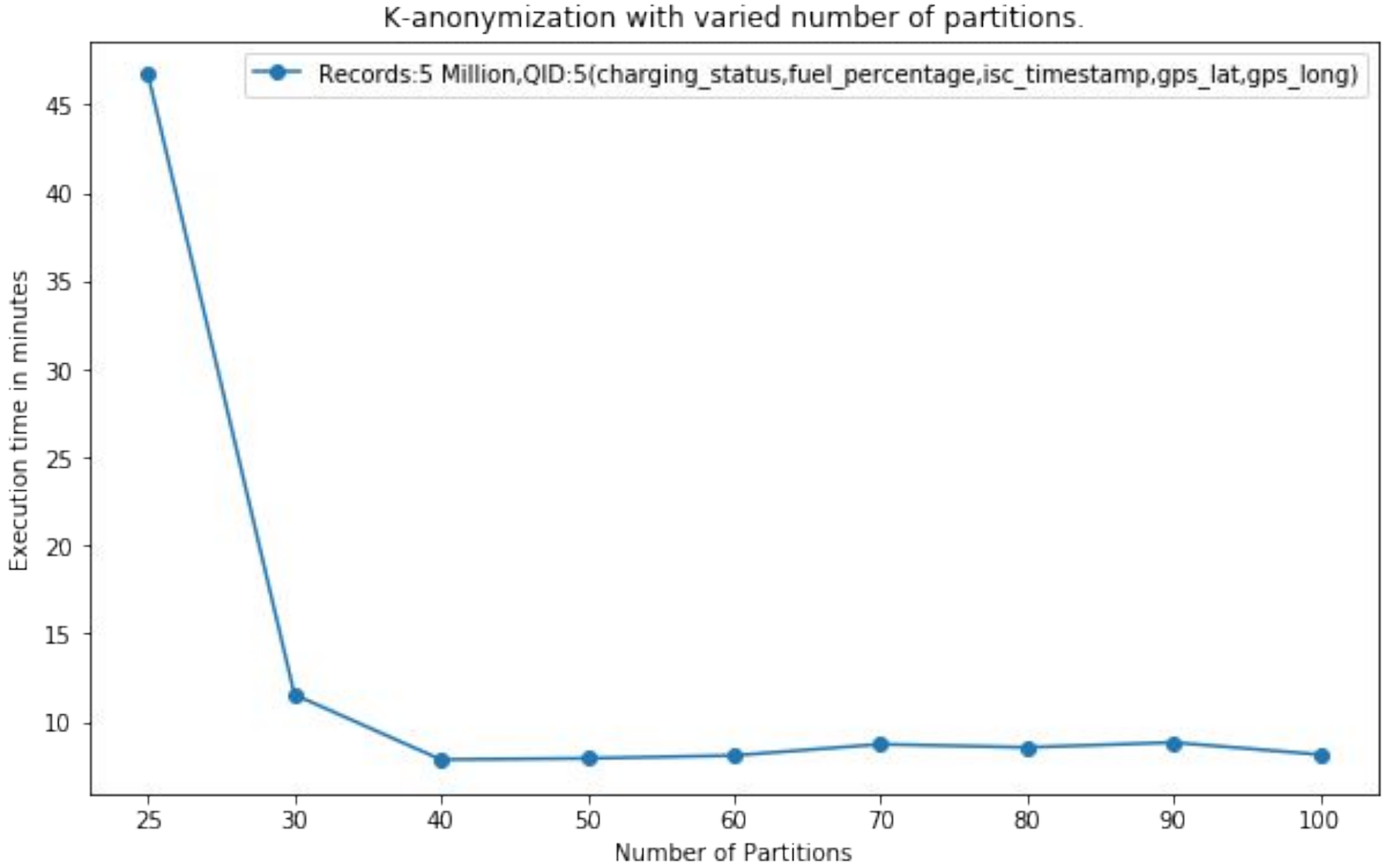
© sebis

Results(4/7): Execution time vs Number of Partitions [1 Million records]



© sebis

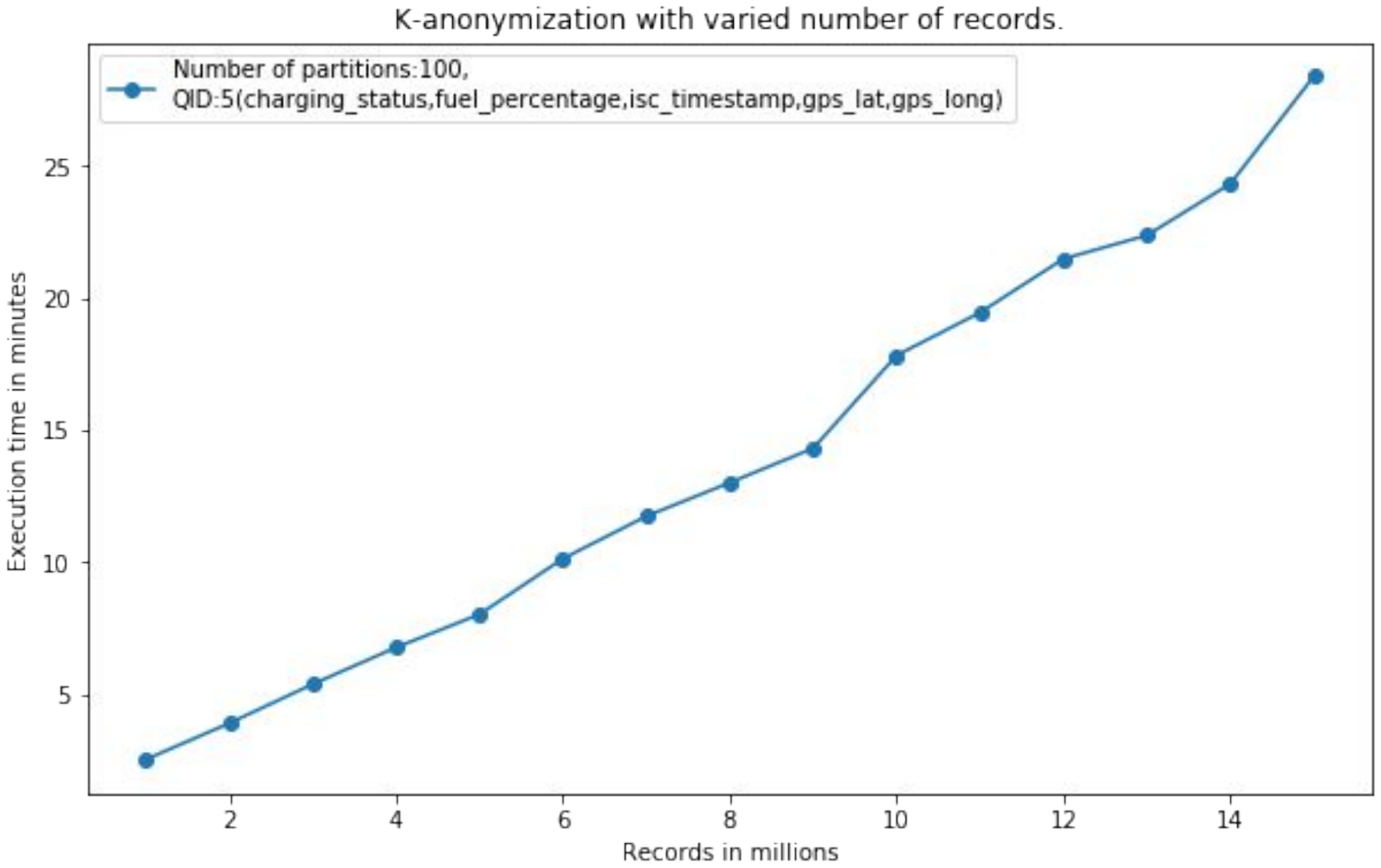
Results(5/7): Execution time vs Number of Partitions [5 Million records]



Out of memory exception for partitions < 25

© sebis

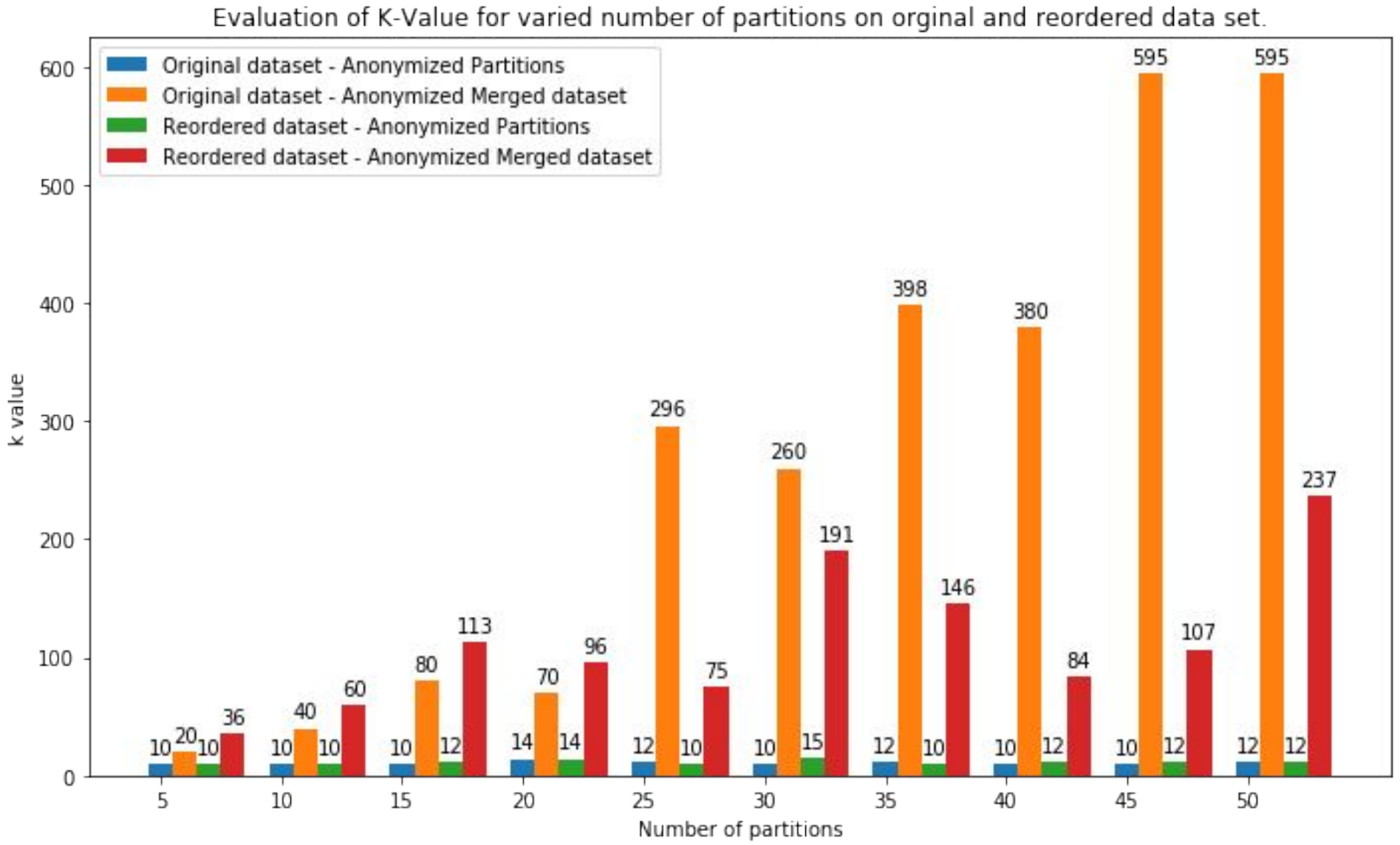
Results(6/7) : Execution time vs Number of records in data set



Out of memory exception for number of records greater than 15 million

© sebis

Results(7/7):K-value of original and reordered data set vs number of partitions.



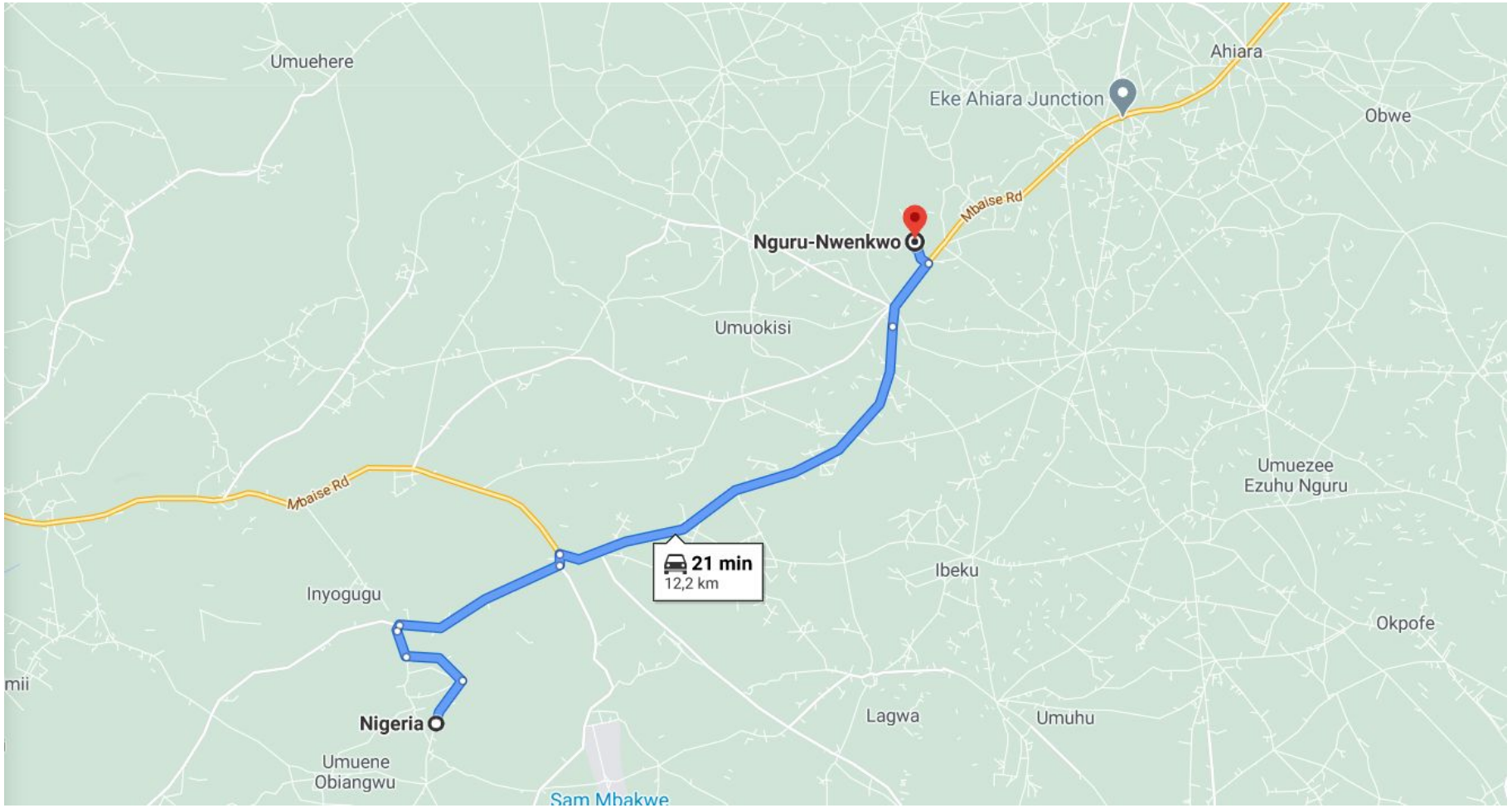
Original dataset : Dataset with no sorted recorded.

Reordered dataset : Dataset with QID fuel_percentage values sorted in ascending order.

© sebis

Insights from anonymized dataset

By performing queries on anonymized data set where fuel_percentage for cars was '*very low*', the range of latitude and longitude - [5.441-5.509][7.181-7.241] were found.



© sebis

Conclusion and Future Work

Conclusion

- Listed properties of the three primary k-anonymization techniques chosen - Mondrian, Incognito and Datafly. A recommendation on the usage of these three algorithms were formulated.
- To apply k-anonymization in distributed manner for relatively large data sets for the use case of Automotive industry, data partitioning using Apache Spark and data anonymization through ARX API was implemented as Proof-Of-Concept.
- From the evaluation, its proven that proposed k-anonymization approach works better than the centralized approach.
    - For relatively medium to small data sets(~size=1M), the execution time was halved.
    - For relatively larger data sets(~size=5M), anonymization could only be applied using the proposed approach..

Future Work

- Solution can be extended by deploying the application in high computing cloud environment which would provide better computational power to process even bigger data set.
- Solution can also be extended to other privacy models like l-diversity, t-closeness, and differential privacy.
- As proposed by Rohde and Schwarz, dataset generator could be enhanced to use concepts of statistical models for each attribute in data set.

© sebis

Sharada Sowmya

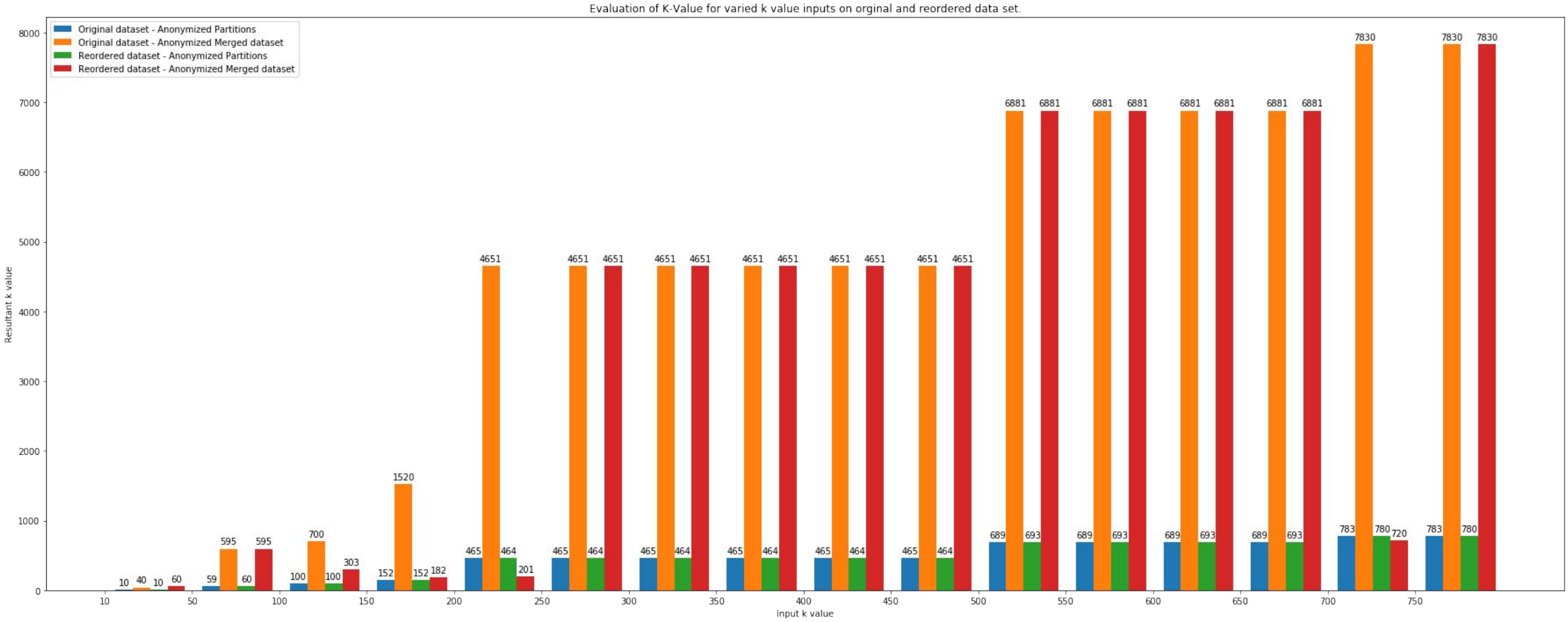
Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289. 17132
Fax +49.89.289.17136

matthes@in.tum.de
wwwmatthes.in.tum.de

Input K-value vs output K-value



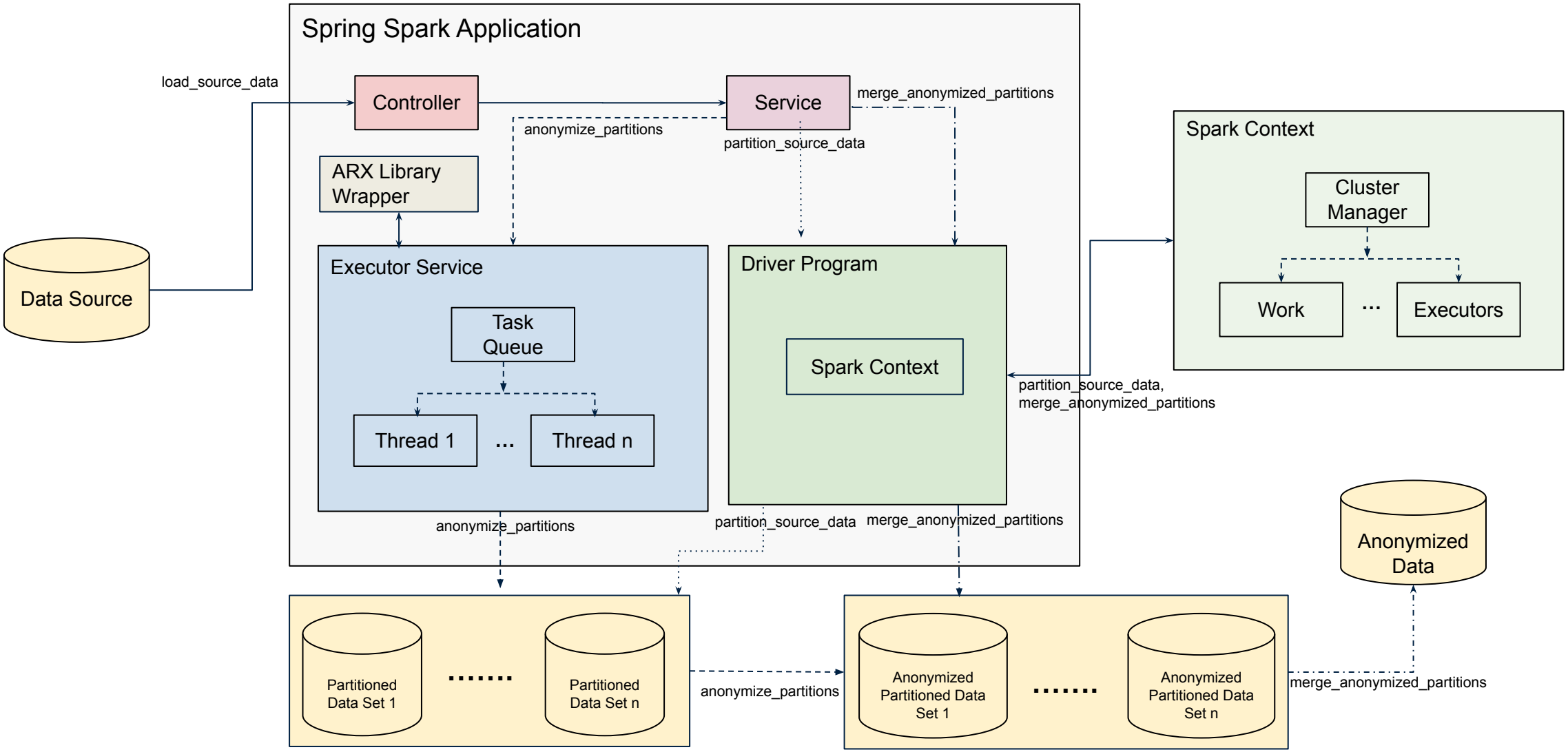
© sebis

Data Preparation

- Created synthetic dataset using the properties of BMW car dataset.
- The primary properties of the data set are as follows:
  - Size of the data set varies from 100,000 to 15 million records.
  - Journeys of up to 1 to 15 cars are synthesized.
  - Car identifiers are generated using custom Universally Unique Identifier(UUID) generator.
  - Charging Status can be CHARGING_ACTIVE or CHARGING_INACTIVE. Once the fuel percentage of the car reaches 0, charging status is changed from CHARGING_INACTIVE to CHARGING_ACTIVE.
  - Fuel Percentage of every car starts with 100% when the car journey is instantiated. This is logically reduced as the car journey progresses.
  - Once the fuel percentage of the car reaches 0%, the car is assumed to be charged completely till fuel percentage is 100% and remaining charging time is 0 minutes.
  - Initial mileage of the car is randomly generated and incremented periodically.
  - Initial timestamp is the randomly generated and incremented periodically.
  - GPS Latitude(gps_lat) and GPS Longitude(gps_long) is captured every 5 miles.
  - The temperature(temperature_external) ranges from -10 to 40 degree Celsius.

© sebis

Software architecture diagram

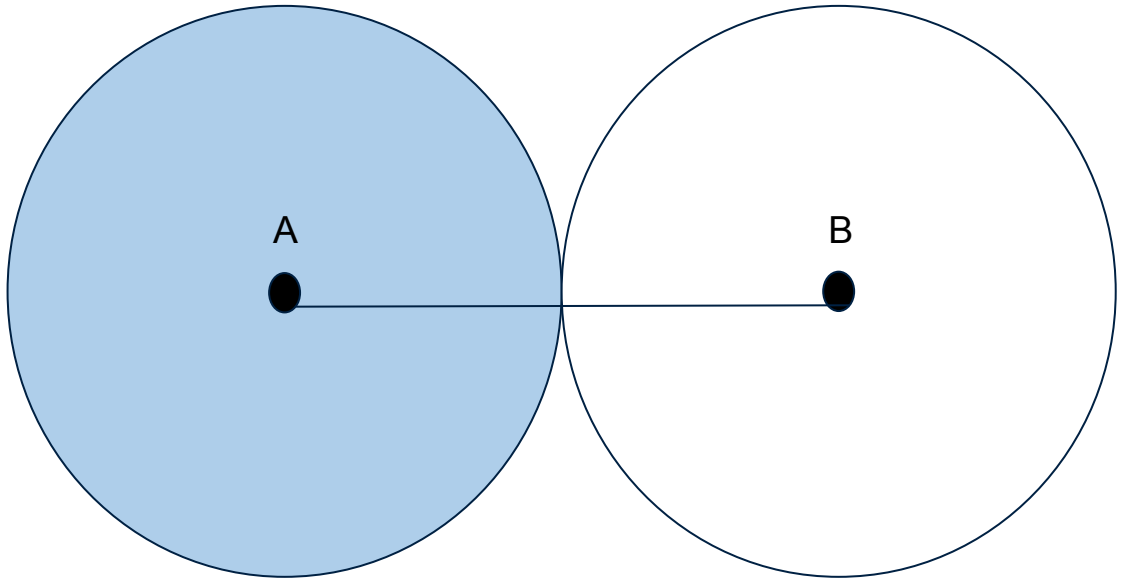


© sebis

Latitude and Longitude

- 0.06 increase in latitude and longitude is 5 miles difference between point.

	gps_lat	gps_long	miles
A	5.479 [5.449-5.509]	7.211 [7.181-7.241]	5
B	5.539 [5.509-5.569]	7.271 [7.241-7.301]	5



- 0.03 decrease will give the lower limit for range
- 0.03 increase will give upper limit for range
- combining lat and long in range - all permutations give all lat and long within the circle

© sebis

ARX - Data Anonymization Tool

ARX is a cross-platform anonymization tool for analyzing and reducing uniqueness of records in relational structured datasets

- ➔ Scalable upto 50 dimensions.
- ➔ Support to multiple privacy models like k-anonymity, t-closeness, l-diversity, δ-Disclosure privacy.
- ➔ Provision to decide data utility measure.
- ➔ Uses combination of Mondrian and Incognito algorithms in a top-down specialization manner.

Pros

- Proven and widely used solution for data anonymization using k-anonymity.
- The solution can be further extended to other privacy models.

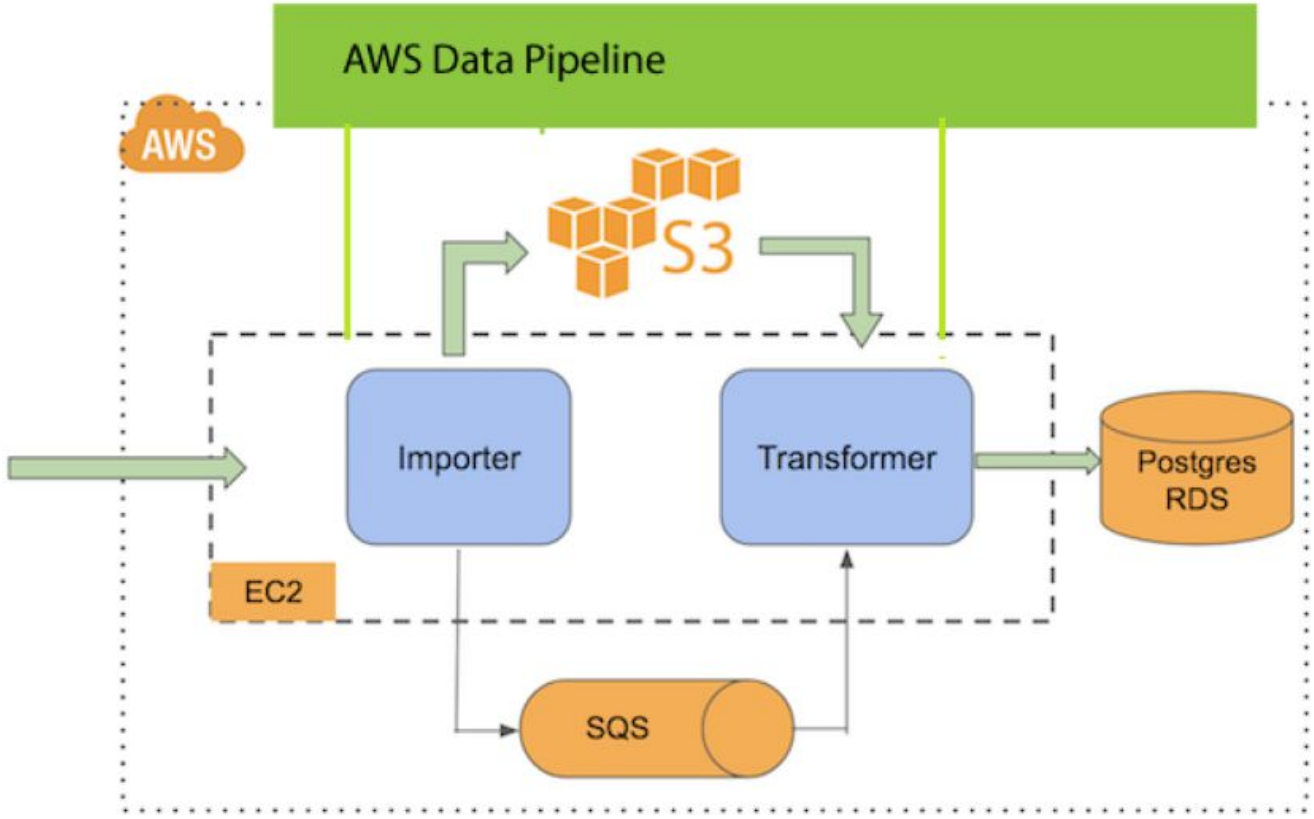
Cons

- No support for anonymization of location data.
- Since the tool uses centralized approach, issues exist with space and time complexities.

© sebis

Software architecture diagram - First version

Framework for a distributed privacy model pipeline.



Step 1: Data Importer is an EC2 instance which will fetch the data from an external API.

Step 2: Data is backed up into S3.

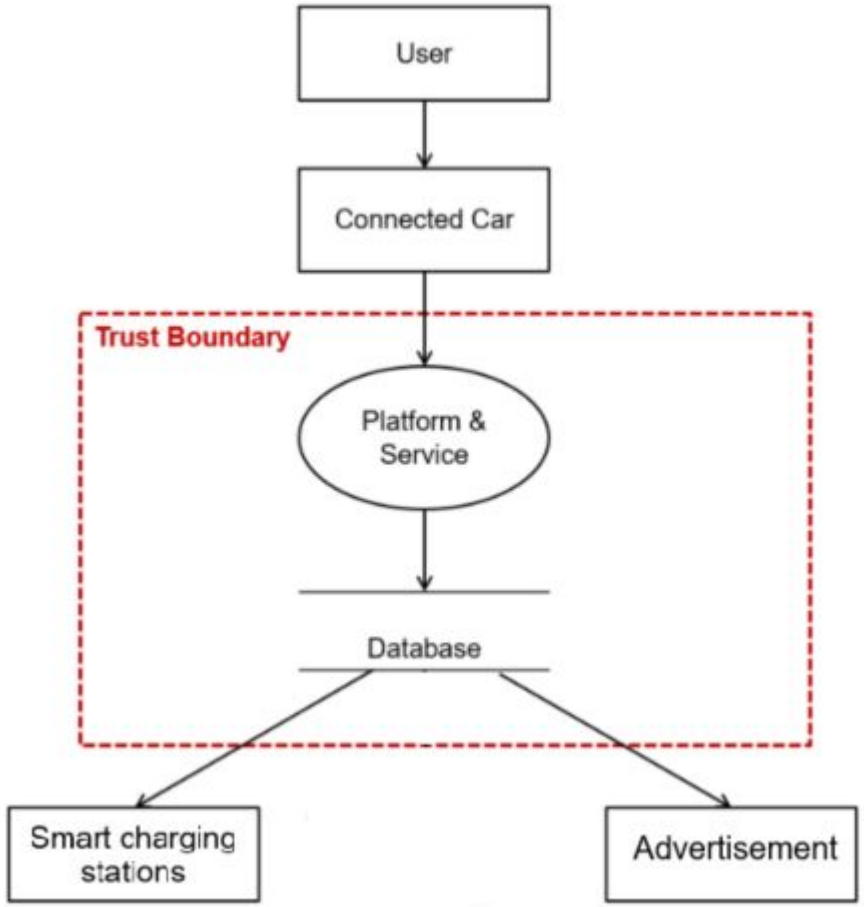
Step 3: Data Importer will send a message to queue on SQS once data backup is complete.

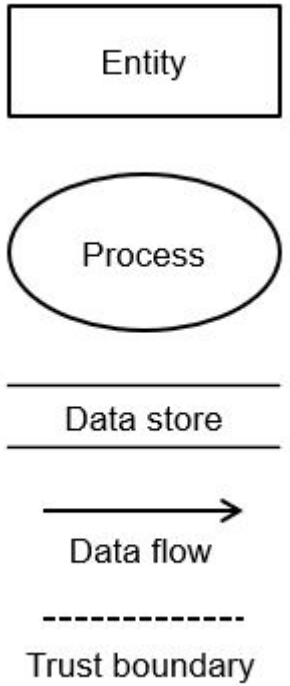
Step 4: Data Transformer is an EC2/EMR instance which listens to the queue and starts de-identification process.

Step 5: De-identification process includes partitioning of the dataset and applying de-identification methods in a distributed manner using AWS Data pipeline.

Step:6: Storing anonymized data into Postgres database is the last stage of the data pipeline.

Data-flow diagram(DFD) of the connected car use case

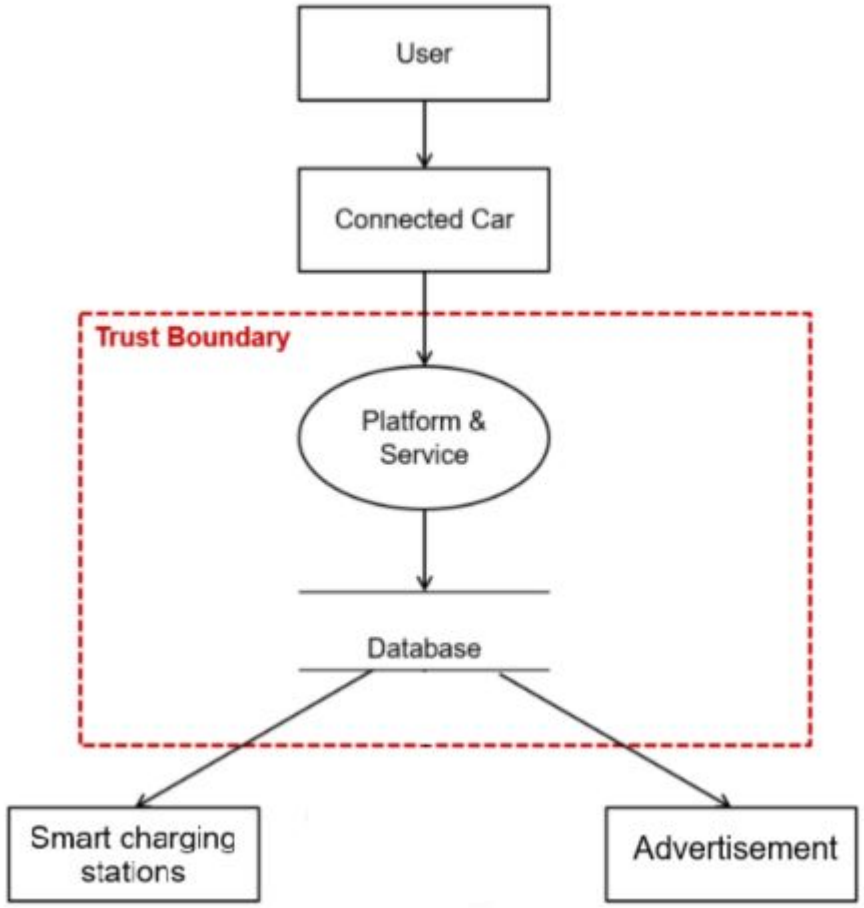




© sebis

Privacy threat modelling of the DFD - LINDDUN methodology

	Threat target
Data Store	Database
Data flow	User data stream (user-connected car)
	Connected car data stream (connected car-platform)
	DB data stream (platform-database)
	Third party data stream (DB-Third party applications)
Process	Platform & Service
Entity	User
	Connected Car
	Third party applications



© sebis

Privacy threat modelling of the DFD

	Threat target	L	I	N	D	D	U	N
Data Store	Database	X	X	X	X	X		
Data flow	User data stream (user-connected car)	X	X	X	X	X		
	Connected car data stream (connected car-platform)	X	X	X	X	X		
	DB data stream (platform-database)	X	X	X	X	X		X
	Third party data stream (DB-Third party applications)	X	X	X	X	X		
Process	Platform & Service	X	X	X	X	X		
Entity	User	X	X				X	
	Connected Car	X	X				X	

Linkability

Identifiability

Non-repudiation

Detectability

Information **D**isclosure

Content **U**nawareness

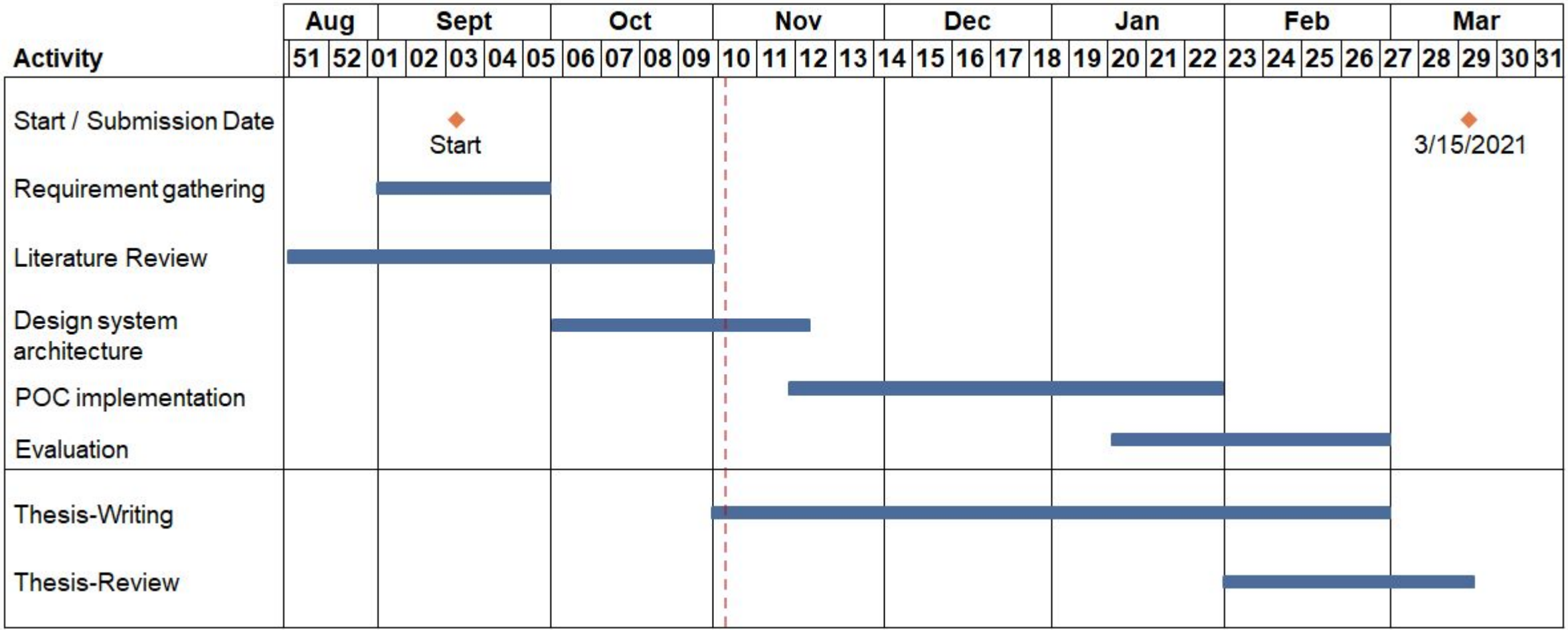
Non-compliance

X Privacy threat mitigation through de-identification is possible.

X Privacy threat mitigation through de-identification is not possible.

© sebis

Timeline



© sebis