Name: **Nurlan Imanov (G49119771)**

July 31, 2023

Title: **Weekly Report 4(Designing database to efficiently store and retrieve massive amounts of data)**

1 Introduction

Since it is my first report regarding the project I would like to start with the general situation of the project. As it is obvious I was not able to start the project because the topic that I selected was too broad. Fortunately, after discussion with colleagues and personal investigation, I came up with an idea and started the project. I know the remaining time is very limited I will try to do my best at least to finish some parts of the project.

2 What is the project about?

As the title suggests the project is about the design of the database for Big Data. The scenario is: The company currently stores its textual data in RDBMS(PostgreSQL). As the data grows daily, they face some performance issues with queries in the current design of the database. Generally, database optimization is needed. Apart from that, the company knows that they will face Big Data issues in the near future. Since the migration of the database is costly it is decided to start with optimization of the current database. For that reason, the project will start by investigating of optimization of RDBMS for Big Data.

The second phase of the project is to test the performance metrics of RDBMS and NoSQL and compare them in terms of handling Big Data. For that, it is planned to conduct load tests in both RDBMS and NoSQL to simulate the big data issues that the company will face in near future.

The main goal of the project is not just to use the databases as storage. It is aimed to investigate the papers that propose methods to design the architecture/design of the database and test the possible real scenarios.

As I work as a Data Engineer currently I am in charge of optimizing the ETL pipelines of the company. This project focuses on the possible database bottleneck of the ETL pipeline.

3 What has been done so far?

Papers related to Big Data were read in order to define what is the Big Data issues to know for which cases it can be considered to have Big Data problems in the system.

As mentioned in the above scenario currently the company stores the textual data in RDBMS and there are some performance issues. Therefore, papers related to database optimization of RDBMS for Big Data are read and it is planned to apply them to the database in the upcoming week.

For that, the consumed time of the queries will be calculated and optimization techniques will be applied to enhance them.

4 What is planned next?

The main goal is to prove with the metrics that NoSQL databases are more appropriate for Big Data. After making some improvements in RDBMS it is planned to move the NoSQL database and conduct the same experiments there. The same approach will be applied to the NoSQL database as well. Papers will be reviewed to enhance the performance of NoSQL for Big Data cases. To simulate the Big Data issues it is planned to make experiments under load.

Learning optimization of databases for Big Data cases is the main objective of the project.