Division of Industrial Electrical Engineering and Automation | LUNDS TEKNISKA HÖGSKOLA | PRESENTED 2020-06-12

THESIS Data utilization and aggregation
STUDENTS Maximillian Vilensten & Oskar Hermansson
SUPERVISOR Christian Nyberg (LTH)
EXAMINER Mats Lilja (LTH)

Employing data lakes to collect and analyse dynamic sensor data

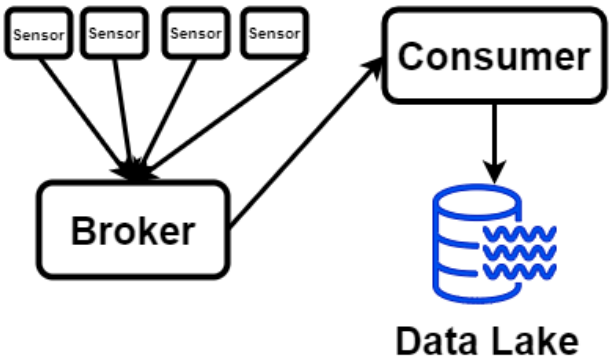
Popular science summary **Maximillian Vilensten & Oskar Hermansson**

The amount of data is steadily increasing and with it comes data from new sources and new formats. The ever changing state of data makes storing data challenging as it most storage options only takes data on a format specified when storing it. This thesis suggests a solution to this for a specific use case.

Today most data is stored with rigid structures, this means that the database expects a specific type of input which then will be sorted and analyzed. This option works well if one knows exactly that the data received maintains the same format for a long period of time. But with different types of data or in situations where the data format might be changing this traditional data storage methodology starts to introduce some issues, since you now have to somehow make sure that all data now has to follow the same format before being sent to the database. Therefore it would be interesting to find out if there is a way to create a database that can store data in all sorts of formats without having to spend time to get them into the right format first. But in also needs to be as accesible and cost effective as the traditional database otherwise it does not benefit the anyone.

In this thesis we have developed a cloud database that can store and analyze data regardless of format. This type of database is called a Data Lake and has grown popular amongst many large companies handling large amounts of data such as Amazon. We also researched the best way to get the data into this database from the outside sources such as movement sensors. To do this a comparison and analysis of different solutions to send data was also made and the winner would then be included in the final prototype.

In order to find out which data storage solution that would be best multiple solutions where tested where the Data Lake on the end was the winner and was therefore used for the final prototype. When it came to finding was to send data to this data storage solution a similar test was held to find out which was best for the data storage we made. In the end a software called Apache Pulsar was the winner and was also used in the final prototype.



In the end a Data Lake was developed that could take a lot of different types of data and still held a high level of adaptability which makes it a possible replacement to a traditional database. The program used for getting data into the Data Lake held up nicely while testing and produced good results in the amount of data it could handle. While still not affecting performance and cost to much in comparison to a traditional database.