The University of Queensland
INFS3208 Cloud Computing

INDIVIDUAL PROJECT PROPOSAL

INTRODUCTION:

This individual project is a Type II Project: Big Data Focused - Big Data Queries, primarily focusing on analysing Queensland flora datasets of 2019 and 2022, assisted by a threatened species table recorded in 2019. Data sets are taken from the Queensland Government Open Data Portal.

Australia, renowned for its biodiversity, faces the challenge of conserving its unique flora amidst rapid environmental changes. However, tracking and preserving the threatened species in Australia became a big issue since the data is significant, and it is hard to manage the preservation progress in each state. Therefore, finding a method that can track and determine the number of threatened species and the trend through the years is essential.

The analysis consists of determining which state has the most endangered species and arranging them in descending order. Therefore, illustrate trends and patterns of those threatened species in each state after three years (2019 - 2022). This computation can provide oversight about the habitats and species in each state, thus helping users to conclude about the preservation progress and find a plan on time.

Traditional computing platforms, while reliable, often need to catch up when grappling with voluminous datasets. Three datasets provide approximately 40,000 records, which consume a lot of effort and labour. Their limited processing capabilities lead to extended computation durations, and they need more scalability and real-time data processing features crucial for large-scale projects.

In contrast, cloud platforms can provide scalability, handle large data sets effortlessly, and carry out computation correctly with incredible speed. Moreover, their inherently distributed architecture ensures swifter data analytics. Within the cloud environment, this project can guarantee the quality of data and computational speed with a promising cost efficiency.

TECHNICAL SOLUTIONS:

1. **MariaDB:** Data storage, contains .sql files that are used to store the data from .csv files into the database.

2. **JupyterNotebook:** Using SparkSQL to manipulate and access the database, this can be accessed through the http://external_ip:8888 from the browser**.** This platform facilitates dynamic data analysis, visualisation, and documentation

3. **Spark Containers:** Including spark master container and two worker containers. The spark master container ensures tasks are allocated and executed correctly. On the other hand, the other two containers will compute the resources and parallel processing data.

The virtual machine's initial expense is approximately $172.89 monthly consists of the N1 series, equipped with 4 vCPUs, 8 GB memory, and a 50 GB balanced persistent disk. Post-discounts, the expenditure dwindles to $123.05. Noting that since the project is not operational perpetually, and the persistent disk usage remains below the 50 GB mark, the actual price may be lower

ARCHITECTURE DESIGN: