PROJECT PROPOSAL

Team-Members:

Pooja Pramod Kantrod - A20482083

Raj Banker- A20470311

Naga Surya Suresh Lnu- A20492550

Rahul Maddula - A20488730

Objective: These are the days of innovation for a better future, and businesses must recognize and accept the importance of Big Data in solving complex and difficult problems through better decision making. The term "Big Data" refers to a collection of large datasets with data sizes in petabytes and higher, with high growth and complexity, making traditional database technologies difficult to analyze. This project discusses the various techniques for processing big data and compares those technologies in terms of efficiency, computing power, ease of use and implementation, and predictive modeling.

Data Set Source: The dataset has been taken from Google's Big query, a platform to pull a large dataset from a serverless warehouse. In this project we are going to use the Stack overflow dataset in which tables are pulled out from the BigQuery in csv format.

Proposed Approach: Tools

- Hadoop
- Hive
- Pig
- Spark
- Google Big Query

References:

1. S. Aravinth, A. Haseenah Begam, S. Shanmugapriyaa, and S. Sowmya, "An Efficient HADOOP Frameworks SQOOP and Ambari for Big Data Processing," IJIRST, vol. 1, no. 10, 2015. [Accessed 10 October 2019].

2. A. Fuad, A. Erwin, and H. Ipung, "Processing performance on Apache Pig, Apache Hive and MySQL cluster," IEEE, no. 10110920147010600, 2019. Available: 10.1109/ICTS.2014.7010600 [Accessed 17 October 2019].