TEAM MEMBERS: Tejasvini Shelar, Vineetha Alluri

OBJECTIVE: To collect data, store it in a data warehouse and perform analytics on it to give us meaningful insights.

DATA SOURCE: <https://catalog.data.gov/dataset/purchase-card-pcard-fiscal-year-2013>

We plan on loading the data into the data lake using stream sets since it looks simple and interesting to explore. Among the different data lakes available we use Hadoop hive and impala as our data lakes because we have experience in SQL and we can also visualize the outputs in the form of graphical representation. We would also like to generate reports using tableau to make us familiar with the environment as well. In the future we can do data modelling and normalize our tables to make it more efficient.

We use streaming to generate a continuous flow of data. HDFS is a primary data storage system used by Hadoop applications. It helps in providing high performance access to our data. Impala is an open source massively processing SQL query engine used for querying data on Hadoop. Hive is used for similar purpose, but impala is faster than hive in executing queries since hive uses map reduce operation which takes a lot of time. Tableau is one of the may business intelligence tools which helps anyone understand the data and also helps in deriving meaningful insights.