# Big Data on AWS– Final End to End Project Specification Document Trainer / Mentor – Venkat

***Abstract*—**

More and more E-commerce Websites provide prod- ucts with different prices which made it hard for consumers to find the products and services they want. In order to overcome this data overload, personalized recommendation engines are used to suggest products and to provide consumers with relevant data to help them decide which products to purchase. Recommendation engines are highly computational and hence ideal for the Hadoop Platform. This system aims at building a book recommendation engine which uses item or user based recommendation from Mahout for recommending books. It will analyze the data and give suggestions based on what similar users did and on the past transaction history of the user.

*Keywords:* Hadoop, Personalization, Recommendation, Websites, HDFS, Pyspark, Sqoop,Aws Glue ,DynamoDB,S3, EMR , Databricks,Tableau

The amount of information exposed to people currently exceeds their ability to consume it, leaving many overpowered by the new content available. Ideally, it is preferred that machines and algorithms help to find more interesting matter to individual preferences so that attention can be easily focused on these items of importance.

For solving the problem of making recommendations cus- tomized for each use, recommendation systems apply knowledge discovery techniques. The scale of recommendation poses new challenges for recommendation systems due to high amount of web traffic. These systems have to face the dual challenge producing high quality recommendations as well as calculating personalized recommendations for millions of users. In recent years, these challenges have been met by collaborative filtering(CF). CF technique analyze the user-item matrix to identify relationships between different users and use that information to develop a customized recommendations for each user[1].

Recommendation Engine can be built by a large scale dis- tributed batch processing infrastructure known as 'HADOOP','SPARK' The distributed processing of large data sets called big data across clusters of computers using PYSPARK and HDFS can be done using Hadoop. It makes it possible to store unstructured data by making use of HDFS and other projects that work with Hadoop like HIVE,DYNAMOD DB, SPARKSQL, and Memory process and many more