CASTEK

CENTER FOR
ADVANCED SKILLS & TECHNOLOGIES

BIG DATA Project (Web VDA)

PROJECT TITLE:

Web VDA (Web Visitor Data Analytics) Using Big Data Ecosystem

PROJECT DESCRIPTION:

Web Visitor Data Analytics project is initiated to demonstrate your capability on Big Data Processing & Analytics.
This project involves analyzing log data of the web visitors coming from web server.
Huge log data is first downloaded & ingested into HDFS - Hadoop Distributed File System using Apache Flume utility. After ingesting, data will be cleaned & transformed using Apache Pig/Hive utilities.
Cleansed structured data is then transferred to a relational database storage system where it can be further used for reporting.

TOOLS & TECHNOLOGIES To Be Used:

Operation System: Ubuntu 12/14.04 Server
Data Ingestion: Apache Flume
Data Transfer: Apache Sqoop
Data Storage: HDFS (Hadoop Distributed File System), MySQL
Data Analysis & Transformation: Apache Pig & Apache Hive

WEBLOG LOCATION ON SERVER:

http://bizmap.in/data/MMC_web.log

ANALYTICS REQUIRED:

1) Which is the most viewed page on the web portal
2) Which is the most viewed products on the portal
3) Which is the most frequently used web browser
4) Generate a report with top 3 viewed products of year 2012 & 2011
5) Generate a report with top 3 IPs address accessing portal in year 2012 & 2011
6) Generate 3 different report showing products accessed by top 3 IPs address, reports should have products names & their view count in descending order
7) Generate a report containing all products & their view counts in descending order
8) Generate a report containing all User IPs & their hit counts in descending order