corpus software
...turning ideas into revenue

# Case Study: Data Analytics for a Social Ad Network

## *Existing Platform*

- Ad publisher network of about 1.5 Million websites
- Using AWS (Amazon Web Services) Public cloud for tracking web pages
- 50+ virtual machines using "nginx" are used for capturing the event logs from the network
- The event logs are processed for analytics and business intelligence.

## *Challenges:*

- Daily Volume: The 1.5million website network generates more than 500Million web impressions and over 1Billion events are tracked from those impressions.
- Speed: The event logs are required to process in real-time and in hourly batch to BI application

## *Solutions Implemented*

### FOR REAL-TIME ANALYTICS

The "nginx" logs were streamed from each log server on to MongoDB and Cassandra for 2 different purposes. The stream data is parsed in Python and C modules and loaded in to MongoDB for real-time analytics for publishers. The analytics interface for publishers is developed on top of MonogoDB using Php and open source graph modules.

**HIGHLIGHTS**

- Implemented MongoDB and Cassandra Database
- The logs from 50+ 'nginx' virtual system are streamed to MongoDB, which are parsed in Python and 'C' and loaded for real=time analytics for publishers
- Graphical BI Application for publisher was developed using PHP & open source graph module, on top of MongoDB
- Cassandra receives URLs from publishers to crawl and extract key taxonomy and store; for advertising and other publisher based intelligence.
- **Achievement:** Real time analytics & business intelligence

The data coming from publishers has URL's that were then sent to Cassandra to crawl and extract key taxonomy and store in Cassandra for advertising and other publisher based

intelligence. The core engineering team at the client developed the rela-time analytics component while Conferrasoft responsibilities were to automate the server infrastructure launch, prepare and install all needed modules, monitoring, troubleshooting and fixing any small to medium complex issues pertaining to the application or MongoDB.

## FOR BATCH ANALYTICS:

Using Flume "nginx" server logs were streamed to Hadoop (HDFS). Once the data reached HDFS it is transformed in to Hive tables and loaded in to Stage DW tables. Using Hourly batch jobs the stage data gets transformed and loaded in to BI data warehouse that lives in MySQL using Sqoop from Hive. The Hive batch jobs run on AWS clusters every hour on a dedicated cluster. Once the data reached in to MySQL DW model, using Pentaho business intelligence reporting generated.

### HIGHLIGHTS

- Implemented Hadoop (HDFS), HIVE and Pentaho
- The logs from 50+ 'nginx' virtual system are streamed to Hadoop using Flume
- Log Data is transformed from HDFS to Hive tables and loaded into Stage Data warehouse system
- Hourly batch jobs are created for loading Log data from data warehouse to BI data warehouse (On MySQL using sqoop)
- Pentaho BI tool is used to generate hourly analytics & business intelligence reports.

**About Corpus:**

Corpus Software is one of the faster growing IT solution and services company focused on Digital Media Entertainment, Embedded technology and Business Analytics with offices and partners across Americas, Europe, APAC, Middle east & Africa. We work with clients in most emerging technology, that's where we make their business strong and bring in real difference in the way peer operates. A diverse workplace with continues focused towards developing unique ideas and contributions to make our clients business grow, and to keep the momentum going.

**Offices:** Dallas, London, Singapore, Johannesburg, Hyderabad and Bangalore