REAL TIME PROJECT:

Click Stream Data Analytics Report Project

ClickStream Data

ClickStream data could be generated from any activity performed by the user over a web application. What could be the user activity over any website? For example, I am logging into Amazon, what are the activities I could perform? In a pattern, I may navigate through some pages; spend some time over certain pages and click on certain things. All these activities, including reaching that particular page or application, clicking, navigating from one page to another and spending time make a set of data. All these will be logged by a web application. This data is known as ClickStream Data. It has a high business value, specific to e-commerce applications and for those who want to understand their users' behavior.

More formally, ClickStream can be defined as data about the links that a user clicked, including the point of time when each one of them were clicked. E-commerce businesses mine and analyse ClickStream data on their own websites. Most of the E-commerce applications have their built-in system, which mines all this information.

ClickStream Analytics

Using the ClickStream data adds a lot of value to businesses, through which they can bring many customers or visitors. It helps them understand whether the application is right, and the application experience of users is good or bad, based on the navigation patterns that people take. They can also predict which page you are most likely to visit next and can-do Ad Targeting as well. With this, they can understand the needs of users and come up with better recommendations. Several other things are possible using the ClickStream Data.

Project Scope

In this project candidates are given with sample click stream data which is taken from a web application in a text file along with problem statements.

- Users information in MySQL database.
- Click stream data in text file generated from Web application.

Each candidate has to come up with high level system architecture design based upon the Hadoop eco systems covered during the course. Each candidate has to table the High-level system architecture along with designed eco systems and pros and cons will be discussed with all the other candidates. Finally, will choose the best possible optimal system design approach for implementation.

Candidates are given instructions to create an oozie work flow with the respective Hadoop Eco systems finalized based on the discussion. Candidates has to submit the project for the given problem statement and this will be validated by the trainer individually before course completion.

ECO System involved in click stream analytics Project
➢ HDFS
➢ Sqoop
➢ Pig
➢ Hive
➢ Oozie

Big Data Hadoop Course Content

Chapter 1: Introduction to Big Data-hadoop

- Overview of Hadoop Ecosystem
- Role of Hadoop in Big Data– Overview of other Big Data Systems
- Who is using Hadoop
- Hadoop integrations into Exiting Software Products
- Current Scenario in Hadoop Ecosystem
- Installation
- Configuration
- Use Cases of Hadoop (HealthCare, Retail, Telecom)

Chapter 2 : HDFS

- Concepts
- Architecture
- Data Flow (File Read , File Write)
- Fault Tolerance
- Shell Commands
- Data Flow Archives
- Coherency -Data Integrity
- Role of Secondary Name Node

Chapter 3 : Mapreduce

- Theory
- Data Flow (Map – Shuffle – Reduce)
- MapRed vs MapReduce APIs
- Programming [Mapper, Reducer, Combiner, Partitioner]
- Writables
- Input Format
- Output format
- Streaming API using python
- Inherent Failure Handling using Speculative Execution
- Magic of Shuffle Phase
- File Formats

- Sequence Files

Chapter 4: Hbase

- Introduction to NoSQL
- CAP Theorem
- Classification of NoSQL
- Hbase and RDBMS
- HBASE and HDFS
- Architecture (Read Path, Write Path, Compactions, Splits)
- Installation
- Configuration
- Role of Zookeeper
- HBase Shell Introduction to Filters
- Row Key Design -What's New in HBase Hands On

Chapter 5 : Hive

- Architecture
- Installation
- Configuration
- Hive vs RDBMS
- Tables
- DDL
- DML
- UDF
- Partitioning
- Bucketing
- Hive functions
- Date functions
- String functions
- Cast function Meta Store
- Joins
- Real-time HQL will be shared along with database migration project

CREDO SYSTEMZ

Chapter 6 : pig

- Architecture
- Installation
- Hive vs Pig
- Pig Latin Syntax
- Data Types
- Functions (Eval, Load/Store, String, Date Time)
- Joins
- UDFs- Performance
- Troubleshooting

- Commonly Used Functions

Chapter 7 : sqoop

- Architecture , Installation, Commands(Import , Hive-Import, EVal, Hbase Import, Import All tables, Export)
- Connectors to Existing DBs and DW

Practicals

- SQOOP to import Real Time Weblogs from application to DB and try to export the same to MySQL

Chapter 8 : kafka

- Kafka introduction
- Data streaming Introduction
- Producer-consumer-topics
- Brokers
- Partitions
- Unix Streaming via kafka

Practicals

Kafka

- Producer and Subscribers setup and publish a topic from Producer to subscriber

Chapter 9 : oozie

- Architecture
- Installation
- Workflow
- Coordinator
- Action (Map reduce, Hive, Pig, Sqoop)
- Introduction to Bundle
- Mail Notifications

Chapter 10: Hadoop 2.0 and spark

- Limitations in Hadoop
- –HDFS Federation
- High Availability in HDFS
- HDFS Snapshots
- Other Improvements in HDFS2
- Introduction to YARN aka MR2
- Limitations in MR1
- Architecture of YARN

- Map Reduce Job Flow in YARN
- Introduction to Stinger Initiative and Tez
- Back Ward Compatibility for Hadoop 1.X
- Spark Fundamentals
- RDD- Sample Scala Program- Spark Streaming

Practicals

- Difference between SPARK1.x and SPARK2.x
- PySpark program to create word count program in pyspark

Chapter 11: Big Data Use cases

- Hadoop
- HDFS architecture and usage
- MapReduce Architecture and real time exercises
- Hadoop Eco systems
- Sqoop - mysql Db Migration
- Hive. -- Deep drive
- Pig - weblog parsing and ETL
- Oozie - Workflow scheduling
- Flume - weblogs ingestion
- No SQL
- HBase
- Apache Kafka
- Pentaho ETL tool integration & working with Hadoop eco system
- Apache SPARK
- Introduction and working with RDD.
- Multi node Setup Guidance
- Hadoop latest version Pros & cons discussion
- Ends with Introduction of Data science.

Chapter 12: Real Time Project

- Getting applications web logs
- Getting user information from my sql via sqoop
- Getting extracted data from Pig script
- Creating Hive SQL Table for querying
- Creating Reports from Hive QL