Prasad Jawale DIGAD 20

BDA Experiment 1

Aim: To study hadoop ecosystem and to demonstrate basic hadoop commands

Theory:

1) Hadoop file system was developed using distributed file system design. It is run on commodity hardware.

2) HDFS stores a large amount of data and provids redundancy as data is stored between multiple systems

Features of HDFS

→ Suitable for distributed storage and processing
→ Provides a command interface to interact with HDFS
→ Built in servers of namenode and datanode help users easily check status of cluster
→ Streaming access to file system data
→ File permissions and authentication

3) Different parts of hadoop ecosystem are

1) Data Managment : Oozie, Chuwka, Flume Zookeper
2) Data access : Hive, Pig, Mahoot, Avro Scoop
3) Processing : Map reduce, YARN
4) Storage : HDFS, HBase

Teacher's Sign.: _____

Conclusion: Basic hadoop commands were studied.

Teacher's Sign.: _____

Prasad Jawale D19AD 20

BDA Experiment 2

Aim: To implement word count using mapreduce and RDBMS operations such as selection, union, projection, joins etc using pyspark.

Theory:

① Map Reduce

→ Mapreduce is a programming model and an associated implementation for processing and generating big datasets with a parallel, distributed algorithm on a cluster

Map → Process the input data into small chunks of data

Reduce → Combination of shuffle and Reduce stage to produce output

② Spark

→ Apache Spark is a cluster computing technology designed for fast computation and is based on Hadoop mapreduce

→ It is designed to cover wide range of workloads such as batch applications, interactive queries and stream processing

③ Pyspark

→ Pyspark allows us to work with Resilient Databases (RDD) in python.

Teacher's Sign.: _____

→ Pyspark combines python's learnability and ease of use with the power of Apache spark to enable processing and analysis of data at any size.

Conclusion : word count program & RDBMS operations have been performed.

Teacher's Sign.: _____

Prasad Jawale D16AD 20

BDA Experiment 3

Aim: Use sqoop to load data from RDBMS (weblog \ transaction data) and analyse it using HIVE \ PIG

Theory:

1) Sqoop is a data transfer tool used for transferring data between HDFS, and relational databases such as MySQL, Oracle or SQL server.
2) Sqoop can handle data transfer in parallel, enhancing the transfer by breaking down the data into smaller chunks and processing them simultaneously.
3) When importing data into Hadoop, sqoop extracts data from the source database using SQL queries, transforms it into a format suitable for HDFS and stores it there.
4) It offers features like incremental imports, allowing users to import only new or updated data since the last import.
5) This is particularly valuable for maintaining data consistency and keeping hadoop datasets up to date with changes in source database
6) It offers customisation features such as delimiters, mapping columns, handling null values etc.

Conclusion: Sqoop was used to load & analysis of data was done using HIVE \ PIG

Teacher's Sign.: _____

Prasad Jowale D16AD 20

BDA Experiment 4

Aim: To study HBase shell and perform CRUD (create, read, update, delete) operations on table

Theory:

1) HBase is a open source distributed NOSQL database designed to handle large volumes of data and provide real time read and write access to that data.

2) It is a part of hadoop ecosystem and is often used for applications that require high speed

3) Features of HBase are
   - → It is linearly scalable across various nodes as well as modularly scalable
   - → Consistent read and writes
   - → Easy to use Java API for client access.
   - → Supports Thrift and REST API for non java front ends which support XML, Protobuf and binary data encoding options

4) Create a table
   - → create 'employee', 'Personal info', 'Professional info'

5) Read
   - → get 'table_name', 'row-key'
   - → scan 'table-name'

Teacher's Sign.: _____

6) update

→ put ('employee', '1', 'Personal info:empId', 30)

7) Delete

→ delete ('table-name', 'row_key')

Conclusion: HBase shell and CRUD operations have been implemented

Teacher's Sign.: _____

Prasad Jawale DIGAD 20

BDA Experiment 5

Aim: Create HIVE database and descriptive based statistics, visualization using HIVE \ PIG

Theory

1) Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarise big data.
2) It provides a high level interface for querying and managing large datasets stored in HDFS
3) Features of Hive
   - → Stores schema in a database & processes data into HDFS
   - → Designed for OLAP
   - → Provides SQL type language for querying called HiveQL or HQL
4) Unlike traditional relational databases, Hive follows a "schema-on-read" approach. Data is stored in HDFS without a predefined schema, and the schema is applied at time of query execution
5) Metastore is used to store metadata about the tables, columns, partitions and respective locations in HDFS

Conclusion: Hive database has been created and analysis & visualisation were performed

Teacher's Sign.: _____

Prasad Jawale D16AD 20

BDA Experiment 6

Aim: To implement any one clustering algorithm

Theory:

1) Clustering is the process of dividing data points into a number of groups such that the data belongs to the same group.

2) Types of clustering models

(1) Connectivity models - The data points closer to each other exhibit more similarity to each other than the data points lying further away. This type of model, the data points can be classified as separate clusters & then aggregating them as the distance decreases

(2) Centroid models - These are iterative clustering algorithms where clusters are decided on the closeness of data point to the centroid of the clusters. Kmeans is an example of the centroid clustering model.

(3) Density models - The data points are assigned to a cluster based on the density of the points of the surrounding region.
Example: DBSCAN, OPTICS

Teacher's Sign.: _____

Date
Page

Conclusion

clustering algorithm was implemented using pyspark and ML libraries.

Teacher's Sign.: _____

Prasad Jowale D16AD 20

BDA Experiment 7

Aim: To study and implement page rank algorithm using pyspark.

Theory

1) Page rank is an algorithm used by Google to rank web pages in search engine results.
2) Web pages are considered as nodes in a directed graph, and hyperlinks between pages represent edges.
3) Page Rank of a webpage depends on the number and quality of links pointing to it.
4) Formula for page rank

$$PR(A) = \frac{1 - d}{N + d \cdot \Sigma\left(\frac{PR(T_i)}{L(T_i)}\right)}$$

N - Total no. of pages

d - damping factor

PR(Ti) - PageRank score of page Ti

L(Ti) - Number of outbound links of page Ti

5) Each page distributes a fraction 'd' of its page rank to pages it links to & '1-d' is distributed evenly among all pages

Conclusion

Page rank algorithm has been implemented using pyspark

Teacher's Sign.: _____

Prasad Jawale DIGAD 20

BDA Experiment 8

Aim: To study Twitter data analysis using Flume

Theory:

1) Apache Flume is a distributed, reliable and scalable tool for effeiciently collecting, aggregating, and transporting large volumes of log data, events or streaming data from various sources to centralised data storage, HDFS or HBase.
2) Flume is designed to collect data from various sources, such as web server logs, social media feeds etc.
3) It uses a distributed architecture, and data collection is performed by agents.
4) Once the data is collected, its placed in a channe where data is temporarily stored before it is forwarded to a sink.
5) Sinks are responsible for delivering data from the channel to final destination, which is often Hadoop HDFS or HBase
6) Data collected by flume can be processed and analyzed using tools like mapReduce or Apache spark

Conclusion:

Apache Flume was used to gather and study Twitter Data

Teacher's Sign.: _____