Under the aegis of Vijayam Educational Trust
CATALYST COLLEGE
(A Unit of CIMAGE Group of Institutions)
Institution approved by Education Department, Government of Bihar, Affiliated to Patliputra University, Patna

Ref: CC/WRSP/23/40/45

Date: 06-Mar-2023

NOTICE

This is to inform all the Students that a workshop on Data-Driven Research: Methodologies for Big Data and Hadoop Development will be organized on 21.03.2023 from 9:30 AM to 5:30 PM in the auditorium of Catalyst College.

The workshop is completely free, and no money will be charged for the Training or Certification.

Interested students are instructed to meet the Activity In-Charge / Class Coordinator for more details and their registration.

By the order of

Principal

Principal
CATALYST COLLEGE
Plot No.- C-16(P) Patliputra Industrial Area
Patliputra, Patna-13

Plot No.C16(P), Patliputra Industrial Area
Patliputra, Patna- 800013
(+91) 7250767676
megha@cimage.in

Date: 21.03.2023

Workshop Title:

Data-Driven Research: Methodologies for Big Data and Hadoop Development

Number of Students Participated: 57

Overview:

This workshop focuses on Data-Driven Research and how to leverage Big Data technologies, particularly Hadoop, to handle, process, and analyze massive datasets. Participants will learn how to apply methodologies for data collection, data processing, and data analysis using Hadoop and its ecosystem of tools. The workshop will combine theoretical concepts with practical hands-on exercises to ensure participants can implement these methodologies effectively in their own research or development projects.

Model 1: Introduction to Big Data and Hadoop

Session 1: Understanding Big Data

- Characteristics of Big Data (Volume, Variety, Velocity, Veracity).
- Challenges of Big Data: How Big Data impacts industries like healthcare, finance, e-commerce, and more.
- Data-Driven Research: How Big Data supports research in diverse fields such as genomics, social sciences, and environmental studies.

Session 2: Introduction to Hadoop Ecosystem

- Hadoop Overview: Understanding the Hadoop Distributed File System (HDFS) and MapReduce framework.
- Key Components of the Hadoop Ecosystem: Overview of:
    - HDFS for distributed storage
    - MapReduce for parallel processing
    - Hive for SQL-like querying on Big Data
    - Pig for data flow scripting
    - HBase for NoSQL storage
    - Spark for in-memory processing
- Hadoop Cluster Setup: Introduction to setting up a basic Hadoop cluster (single-node or multi-node).

Session 3: Setting Up Your Hadoop Environment

- Installing Hadoop: Step-by-step guide to installing Hadoop locally or using cloud-based services like AWS, Google Cloud, or Azure.
- Running Basic Hadoop Commands: How to interact with Hadoop using command-line tools.
- Exploring HDFS: Learn how to upload, retrieve, and manage data on the Hadoop Distributed File System.

Model 2: Data Collection, Preparation, and Storage for Big Data Research

Session 1: Data Collection for Big Data Research

- Data Sources: Identifying and collecting data from diverse sources, including IoT devices, social media, public datasets, sensor networks, and enterprise systems.
- Data Formats: Understanding structured, semi-structured, and unstructured data.
- Data Ingestion Tools: Introduction to tools like Apache Flume and Apache Kafka for collecting and ingesting streaming data.

Session 2: Data Storage in Hadoop Ecosystem

- HDFS: Understanding HDFS architecture and how it stores vast amounts of data across multiple nodes.
- Data Partitioning and Replication: How Hadoop ensures data availability and fault tolerance using replication and partitioning strategies.
- Data Security: Discussing data security and access controls in Hadoop environments (Kerberos, ACLs).

Session 3: Data Cleaning and Transformation

- Data Preprocessing: Techniques for cleaning and preprocessing data before analysis.
- Using Apache Hive: Introduction to SQL-like queries in Hadoop with Hive for data transformation.
- Using Apache Pig: A data flow language for processing and transforming data in a high-level way.

Model 3: Hadoop for Data Analysis and Research Methodologies

Session 1: Research Methodologies for Big Data Analysis

- Quantitative vs Qualitative Research: Applying traditional research methodologies in the context of Big Data.
- Exploratory Data Analysis (EDA): Techniques for summarizing and visualizing large datasets.
- Hypothesis Testing with Big Data: Formulating and testing hypotheses using Big Data techniques.

Session 2: MapReduce for Data Processing

- Understanding MapReduce: How MapReduce processes data in parallel across a Hadoop cluster.
- Creating Your First MapReduce Job: Writing a basic MapReduce program in Java (or Python) to process large datasets.
- Optimization: Best practices for optimizing MapReduce jobs for performance and efficiency.

Session 3: Using Apache Spark for Advanced Data Processing

- Apache Spark Overview: Introduction to Spark as a fast, in-memory data processing engine for Big Data.
- Spark SQL: Using Spark SQL for querying structured data.
- Machine Learning with Spark MLlib: An introduction to using Spark for machine learning tasks, such as clustering, regression, and classification.

Model 4: Advanced Data Analysis Techniques and Case Studies

Session 1: Advanced Data Analysis Techniques

- Big Data Analytics Frameworks: Exploring advanced techniques like natural language processing (NLP), graph analytics, and time-series analysis.
- Predictive Analytics with Hadoop: How to use Hadoop for predictive modeling and forecasting.
- Anomaly Detection: Using Big Data to identify unusual patterns or outliers in massive datasets.

Session 2: Using Hadoop for Complex Data Queries

- HQL (Hive Query Language): Deep dive into querying Big Data using Hive.
- Optimizing Hadoop Queries: Techniques for optimizing performance in Hadoop query engines.
- Real-Time Analytics with Apache Storm: Introduction to real-time Big Data processing with Apache Storm.

Session 3: Case Studies: Real-World Big Data Research Projects

- Case Study 1: Big Data in Healthcare – Using Hadoop to analyze medical records and predict patient outcomes.
- Case Study 2: Big Data in Social Media – Analyzing sentiment and user behavior using Hadoop and Spark.
- Case Study 3: Big Data in Retail – Using Hadoop to analyze customer behavior and optimize inventory management.

Model 5: Big Data Project Development and Best Practices

Session 1: Building a Big Data Research Project

- Project Planning: How to define the problem, collect the necessary data, and design your Big Data architecture.
- Choosing the Right Tools: Selecting the right Hadoop ecosystem tools based on project requirements.
- Creating a Research Pipeline: Building an end-to-end pipeline for Big Data analysis from data ingestion to final insights.

Session 2: Deployment and Scaling Hadoop

- Deploying Hadoop in Production: Moving from a development environment to a production environment.
- Scaling Hadoop Clusters: How to scale a Hadoop cluster to handle massive datasets.
- Monitoring and Tuning Hadoop: Using tools like Ambari or Cloudera Manager to monitor, manage, and optimize Hadoop clusters.

Session 3: Final Q&A, Best Practices, and Closing

- Best Practices for Big Data Research: Key takeaways for success in Big Data research using Hadoop.
- Troubleshooting and Debugging Hadoop: Common issues and solutions in Big Data projects.
- Q&A and Wrap-Up: Open session for addressing remaining questions, sharing additional resources, and discussing next steps.

Key Takeaways:

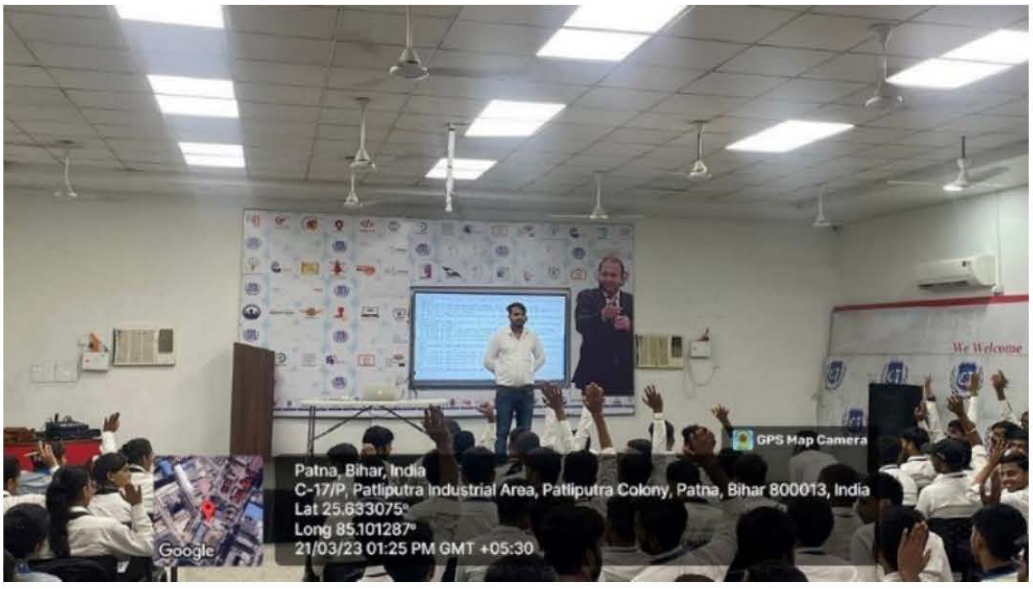
- A solid understanding of Big Data research methodologies and how to apply them with Hadoop.
- Hands-on experience with tools like MapReduce, Hive, Pig, Spark, and more.

- Practical knowledge of data preprocessing, analysis, and advanced analytics techniques.
- Insight into real-world Big Data case studies and how Hadoop is used across various industries for research and analytics.

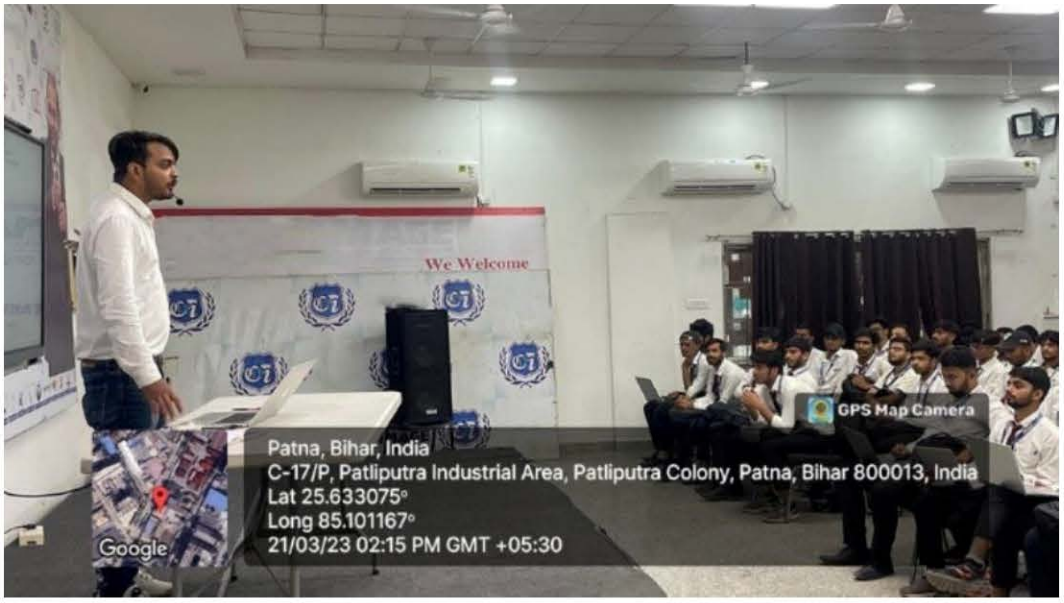
Data-Driven Research: Methodologies for Big Data and

Date:-21/03/2023

Principal
CATALYST COLLEGE
Plot No.- C-16(P) Patliputra Industrial Area
Patliputra, Patna-13



Principal
CATALYST COLLEGE
Plot No.- C-16(P) Patliputra Industrial Area
Patliputra, Patna-13



Data-Driven Research: Methodologies for Big Data and

Date:-21/03/2023

Principal
CATALYST COLLEGE
Plot No.- C-16(P) Patliputra Industrial Area
Patliputra, Patna-13

Registration

For Workshops/Seminars/Conferences during Academic Year 2022-2023

Data-Driven Research: Methodologies for Big Data and Hadoop Development

(21 March 2023)

S. No.	ID	Name of the student	Student's Signature
1	445-9156	Rani Kumari	Rani
2	445-9147	Shubham Kumar	Shubham
3	445-9175	Sonu Yadav	Sonu
4	445-9144	Lavkush Kumar	Lavkush Kr.
5	445-9149	Vikram Kumar	Vikram Kr.
6	445-9162	Ravi Kumar	Ravi Kr.
7	445-9151	Piyush Raj	Piyush Raj.
8	445-9137	Sarika Kumari	Sarika.
9	445-9158	Vikash Kumar	M. Kumar.
10	445-9752	Ayush Verma	Ayush
11	445-9756	Anjali Kumari	Anjali Kri
12	445-9763	Harshit Kumar	Harshit.
13	445-9789	Priyanshu Singh	Priyanshu Singh.
14	445-9792	Rishikesh Kumar	Rishikesh Kr.
15	445-9806	Shalini Mishra	Shalini
16	445-9828	Vivek Kumar	Vivek Kr.
17	445-9831	Nitish Kumar	Nitish Kr.
18	445-9834	Prem Prakash	prem
19	445-9787	Prince Kumar	Prince.
20	445-9849	Nishant Kumar Sumant	N. K. Sumant
21	445-9867	Ankit Raj	Ankit Raj
22	445-9888	Shreya Ranjan	Shreya Ranjan.
23	445-9902	Prashant Kumar	Prashant Kr.
24	445-9908	Ishmeet Kaur	Ishmeet Kaur.
25	445-9931	Rishav Raj	Rishav Raj
26	445-9933	Jaiki Kumar	Jaiki Kumar
27	445-9936	Ritesh Kumar Singh	R. K. Singh.
28	445-9809	Rakesh Kumar	Rakesh Kumar
29	445-9738	Rahul Kumar	Rahul Kr.
30	445-9874	Baibhav Kumar	Baibhav Kumar.
31	445-9019	Golu Kumar	Golu Kr.
32	445-9974	Ashish Kumar	Ashish Kumar

CATALYST COLLEGE

33	445-9886	Muskan Pandey	Muskan
34	445-10019	Deepak Kumar Singh	D. K. Singh.
35	445-9914	Amit Kumar	Amit
36	445-9774	Vishal Kumar	Vishal Kr.
37	445-9855	Aditya Kumar	Aditya Kr.
38	445-9777	Alok Ranjan	Alok
39	445-9782	Shubham Kumar	Shubham Kr.
40	445-9713	Ashutosh Kumar Prasad	A. K. Prasad.
41	445-9905	Gaurav Kumar	Gaurav kuma
42	445-9926	Vivek Kumar	Vivek Kr.
43	445-9839	Shashikant Kumar	S. Kumar
44	445-9917	Sumit Kumar	Sumit kuma
45	445-9836	Sakshi Kumari	Sakshi
46	445-9852	Vishal Kumar	Vishal
47	445-9769	Rahul Raj	Rahul
48	445-9759	Rajesh Kumar	Rajesh
49	445-9726	Kuldeep Kumar	Kuldeep
50	445-9766	Ayush Kumar	A. Kr.
51	445-9881	Sanjay Kumar	Sanjay
52	445-9826	Saurabh Kumar	Saurabh Kr.
53	445-9715	Munna Kumar	Munnu kuma
54	445-9920	Aditi Singh	Aditi Singh.
55	445-9817	Anmol Kumar Yadav	Anmol
56	445-9795	Khalid Ansari	Khalid Ansari
57	445-9732	Chanchal Kumar	Chanchal

CATALYST COLLEGE

(Sign.)

Course Coordinator