ANIL AILANI SIR-9819704277
OFFICE-022-66730730/9324125511

Hadoop Essentials (Effective June 2016)

Introduction to Big Data

- Motivation and Basics
- What is Big Data
- Big Data Opportunities
- Big Data Challenges
- Characteristics of Big Data
- Data Storage and Analysis

Introduction to Hadoop

- A brief History of Hadoop
- Apache Hadoop and Hadoop Eco-System
- Hadoop Architecture
- Hadoop Configuration Types
    - Single Node (Standalone)
    - Pseudo Distributed Mode
    - Fully Distributed Mode (Clusters)

The Hadoop Distributed File System (HDFS)

- The Design of HDFS
- HDFS Concepts
    - Blocks
    - NameNode, SecondaryNameNode and DataNode
    - HDFS Federation
    - HDFS High-Availability

- Hadoop DFS – the Command Line Interface
- Basic File System Operations

Introduction to MapReduce

- Map and Reduce Basics
- Functional Programming Basics
- How MapReduce works
  - Anatomy of a MapReduce Job Run
  - Legacy Architecture
    - Job Submission
    - Job Initialization
    - Task Assignment
    - Task Execution
    - Progress and Status Updates
    - Job Completion and Failures
- The MapReduce Web UI
  - The NameNode Page
  - The JobTracker Page
- Shuffling and Sorting
- Splits, Partition, and Combiner Functions
- Optimization Techniques
  - Speculative Execution
  - Task JVM Reuse
- Types of Schedulers and Counters
- MapReduce 2 (YARN)
  - YARN (Yet Another Resource Negotiator)
  - MapReduce (Classic) vs. MapReduce 2 – Code and Architectural Level
  - Distributed Cache
  - Map Side Join with Distributed Cache

MapReduce Job Types and Formats

- Job Types
  - The Default MapReduce Job
  - The Default Streaming Job
- Input Formats
  - Input Splits and Records
  - Text Input

- Output Formats
    - Text Output

MapReduce Features

- Counters
    - Built-in Counters
    - User-Defined Java Counters
- Sorting
    - Preparation
    - Partial Sort
    - Total Sort
    - Secondary Sort
- Joins
    - Map-Side Joins
    - Reduce-Side Joins
- Brief introduction to Pig and Spark with one example

Note: Example applications for concept and features.

Projects:

The following projects will be part of the curriculum. The **case-studies** of all these projects will be based on **live data** from the industry. Analytics of all output from the Hadoop jobs will be rendered in a graphical form using the JFreeChart library.

Project #1: National Climate Data Center (NCDC) Weather Condition Analysis
Industry: Meteorology
Data: NCDC is the world's largest active archive of weather data. NCDC produces numerous climate publications and responds to data requests from all over the world. NCDC provides access to daily data from the U.S. Climate Reference Network / U.S. Regional Climate Reference Network (USCRN/USRCRN) via anonymous ftp at: ftp://ftp.ncdc.noaa.gov/pub/data/uscrn/products/daily01
or from http://www.ncdc.noaa.gov/
Problem Statement: Analyze the data in Hadoop to:
1. Find the maximum, minimum temperature month-wise and year-wise.
2. Analyze temperature variation for a series of years
3. Analyze error-level of sensor data
4. Determine Fog/Dew and analyze visibility distance
5. Analyze atmospheric pressure

Project #2: Analyze social bookmarking sites to find insights and trends
Industry: Social Media

Data: It comprises of the information gathered from sites like reddit.com, stumbleupon.com etc. which are bookmarking sites and allow you to bookmark, review, rate, search various links on any topic. A bookmarking site allows you to bookmark, review, rate, search various links on any topic. The data is in XML or CSV format and contains various links/posts URL, categories defining it and the ratings linked with it.

Problem Statement: Analyze the data in Hadoop Eco-system to:

1. Fetch the data into Hadoop Distributed File System and analyze it with the help of MapReduce, Pig and Hive to find the top rated links based on the user comments, likes etc.
2. Using MapReduce convert the semi-structured format (XML data) into structured format and categorize the user rating as positive and negative for each of the thousand links.
3. Push the output HDFS and then feed it into PIG, which splits the data into two parts: Category data and Ratings data.
4. Write a Hive Query to analyze the data further and push the output into relational database (RDBMS) using Sqoop.

Project #3: Tourism Data Analysis

Industry: Tourism

Data: The dataset comprises attributes like: City pair (Combination of from and to), Adults traveling, Seniors traveling, Children traveling, Air booking price, Car booking price, etc.

Problem Statement: Find the following insights from the data:

1. Top 20 destinations people travel most. Based on given data we can find the most popular destinations where people travel frequently, based on the specific initial number of trips booked for a particular destination
2. Top 20 locations from where most of the trips start based on booked trip count
3. Top 20 high air-revenue destinations i.e which 20 cities generates high airline revenues for travel, so that the discount offers can be given to attract more bookings for these destinations.

Project #4: Airline Data Analysis

Industry: Aviation

Data: Publicly available dataset which contains the flight details of various airlines like : Airport id, Name of the airport, Main city served by airport, Country or territory where airport is located, Code of Airport, Decimal degrees, Hours offset from UTC, Timezone, etc.

Problem Statement: Analyze the airlines data to:

1. Find list of Airports operating in the Country
2. Find the list of Airlines having zero stops
3. List of Airlines operating with code share
4. Which country (or) territory has the highest number of Airports
5. Find the list of Active Airlines in the United States.

Project #5: Analyze Movie Ratings

Industry: Media

Data: Publicly available data from sites like rotten tomatoes, imdb, etc.

Problem Statement: Analyze the movie ratings by different users to:

1. Get the user who has rated the most number of movies
2. Get the user who has rated the least number of movies
3. Get the count of total number of movies rated by user belonging to a specific occupation
4. Get the number of underage users.

Project #6: Analyze YouTube data
Industry: Social Media
Data: It is about the YouTube videos and contains attributes like, VideoID, Uploader, Age, Category, Length, views, ratings, comments, etc.
Problem Statement: Find out the top 5 categories in which the most number of videos are uploaded, the top 10 rated videos, the top 10 most viewed videos.