Faculty of Computer& Informatics.
Department of Artificial Intelligence &
Data science
Department of Software Engineering

Analyzing Data using Big Data System Techniques

Prepared by

Abdullah Khadem Aljame

Ali Hussain Alaswad

Supervised by

Dr.Eng.Mouhib Alnoukari

Eng.Anas Abdulaziz

Academic Year

2023-2024

Abstract

Content list

Chapter 1: Project Summary

1.1. Project Summary:

1.1.1. Overview:

Big data refers to the large, diverse sets of information that grow at ever-increasing rates. It encompasses the volume of information, the velocity at which it is created and collected, and the variety of the data points being covered.

1.1.2. Aim:

In this project, we are supposed to build a working environment for a big data system and perform data analysis using big data technology.

1.2. Project Goals:

- Build A Real Big Data System In SPU
- Analyzing Data using Big Data Techniques
- Make the environment ready to receive future big data projects

Chapter 2: Weekly Project Plan

Week 1 & 2:

- Data center and Information about it
- Servers types and about the specification
- Chose ubuntu server as the operating system for Hadoop and config it.
  - steps:
    - install ubuntu server at each server
    - configure Server by passing the IP Address,

    Faced Some problem configuration:

    FIRST

    - The operating system did not recognize the server's network card

    SECOND

    - the SSH problem: the way to identify the server with the college laboratories and halls so that we communicate with the servers from the halls

- How to solve all these problems?

- Changing the operating system to Ubuntu desktop22.04

- configure the IP Address for each server

  Master server: spu@10.0.1.219

  Slave server: spu@10.0.1.220

-Now we can access the server from all halls and laboratories using ssh by using " ssh spu@10.0.1.219/220 "

Week 3 & 4:

1-Install jdk 8 for Hadoop nodes (Master and Slave)

2-Install Apache Hadoop 3.3.6 for each server and config it by edit the xml files

Steps:

- install java 8 on each server

-download hadoop 3.3.6 https://www-eu.apache.org/dist/hadoop... and extract

-copy hadoop-3.3.6.tar.gz folder to each slave server with scp and extract

-add hostname in /etc/hosts/

-move folder: hadoop-3.3.6 to /usr/local/hadoop/

-add PATH /etc/environment

JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

-create user: hadoop

-create ssh keygen from master server and copy to slave server

-Configuring the Hadoop Master

-Copy the configuration files to each of your Hadoop Nodes from your Hadoop Master.

-Format the HDFS file system on master server

-Now you can start HDFS

- open master server: http://192.168.56.101:9870

-config Yarn

Week 5 & 6:

Gather, Find and understand the data:

https://www.kaggle.com/datasets/debashis74017/stock-market-data-nifty-50-stocks-1-min-data

about data:

The Nifty 100 index tracks the performance of the top 100 companies listed on the National Stock Exchange (NSE) of India. This broader index includes constituents of the Nifty 50 as well as an additional 50 companies, providing a more comprehensive view of the Indian stock market. The Nifty 100 index offers exposure to a diversified set of stocks across various sectors, making it a valuable benchmark for investors and fund managers. It serves as a barometer of the Indian equity market, allowing for a broader assessment of market performance and trends. The inclusion of additional companies in the Nifty 100 offers investors a more extensive representation of the Indian stock market compared to the Nifty 50, allowing for a broader perspective on market dynamics and diversification opportunities.