Page No.
Date

Assignment 1

* Problem Statement

Develop a mapreduce program to calculate the frequency of a given word in a given file.

* Software Requirements

Jupyter

* Hardware Requirements

500 GB HDD, 2 GB RAM, PIV

* learning Objectives

Understand the MapReduce paradigm & implement MapReduce program to calculate frequency of given word in given file.

* & Prerequisite

Python

* Theory

MapReduce is a Java-based, distributed execution framework within the Apache Hadoop ecosystem. It involves two processing steps that developers

implement i) map & 2) Reduce

Map

In the map phase, input data is divided int blocks & processed by multiple mapper nodes each applying the map function to local data.

Shuffle, combine & partition

During this phase, worker nodes redistribute based on output keys, ensuring all data wit same key is on one node. Optionally, a combi further reduce data on each mapper node. Partitioning assigns data to specific reducers

Reduce

Worker nodes process groups of key-value pairs in parallel, aggregating values for eac Unlike mapping, reducing is optional & focu further processing aggregated data.

Conclusion

Thus developed a mapReduce program to calculate frequency of given word in gi file.

Mapper

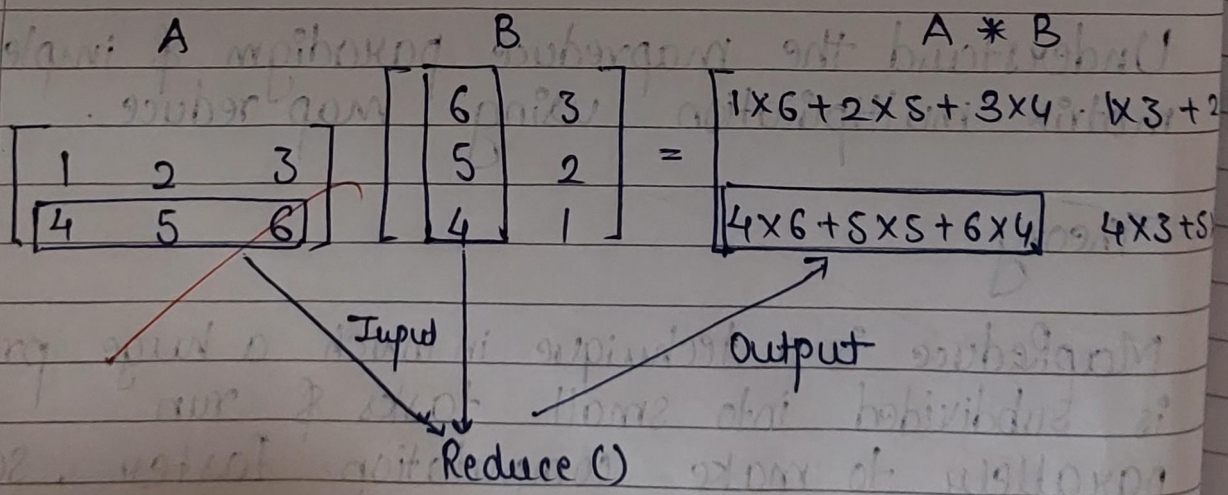
It takes raw data input & organizes into key value pairs.

For example, In a dictionary, you search for word 'Data' & its associated meaning is "facts & statistics collected together for reference or analysis". Here the key is data & the value associated with is facts & statistics collected together for reference or analysis.

Reducer

It is responsible for processing data in parallel & produce final output.

example



Assignment 5

* Problem Statement

Hive: Introduction creation of database & table, Hive partition, Hive built in function & operators, Hive View & Index.

* Requirements

VMWare, XAMPP Server, Web browser, 1 GB RAM, Hard disk 80 GB.

* Learning Objective

learn to create database, hive partition. explore Hive built in functions & operators, Hive view & Index.

* Theory

Hive is an open source data warehousing solution built on top of hadoop. It supports an SQL-like query language called HiveSQL. These queries are compiled into MapReduce jobs that are executed on Hadoop.

While Hive uses hadoop for executing queries, it reduces the effort that goes into writing & maintaining MapReduce jobs.

Hive supports database concepts like tab columns, rows & partitions. Both primitiv (inter, float, string) & complex data typ (map, list, struct) are supported.

Moreover, these types can be composed to support structures of arbitrary complexity. The tables are serialized / deserialized using default serializer / deserializer. Any new data format & typ can be supported by implementing SerDe & Object Inspector java interface

* Conclusion

Thus we have ~~implemed~~ implemented Hi

Assignment 1

* Problem Statement

Import data from different sources such as (excel, SQL server, Oracle, etc) & load in targeted System.

* Title

Import data from different sources such as (excel, SQL server, oracle, etc) & load in targeted System.

* Prerequisite

Python

* Software Requirement

Power BI Tool

* Hardware Requirement

PIV, 2 GB RAM, 500 GB HDD

* Learning Objectives

Page No.
Date

Learn to import data from different sources such as (excel, oracle, etc) & load in targeted system.

* Outcomes

After completion of this assignment students are able to understand how to import data from different sources & load in targeted system.

* Theory

legacy data, according to business dictionary, is "information maintained in an old or out-of-date format or computer system that is consequently challenging to access or handle".

Importing Excel Data

(1) launch Power BI Desktop
(2) From the Home ribbon, select get data. Excel is one of the most common data connections, so you can select it directly from the get data menu.
(3) If you select the get data button directly, you can also select file > excel & select connect.
(4) In the open file dialog box, select the products.xlsx file.
(5) In the navigator pane, select the products table & then select edit.

Page No.
Date

Assignment 2

* Title

Data Visualization from extraction Transformation & loading (ETL) process.

* Problem Statement

Data Visualization from extraction Transformation & loading (ETL) process.

* Prerequisite

Python

* Software Requirements

Jupyter

* Hardware Requirements

PIV, 2 GB RAM, 500 GB HDD

* Outcomes

After completion of this assignment students are able to understand how data visualization is done through extraction Transformation & loading process

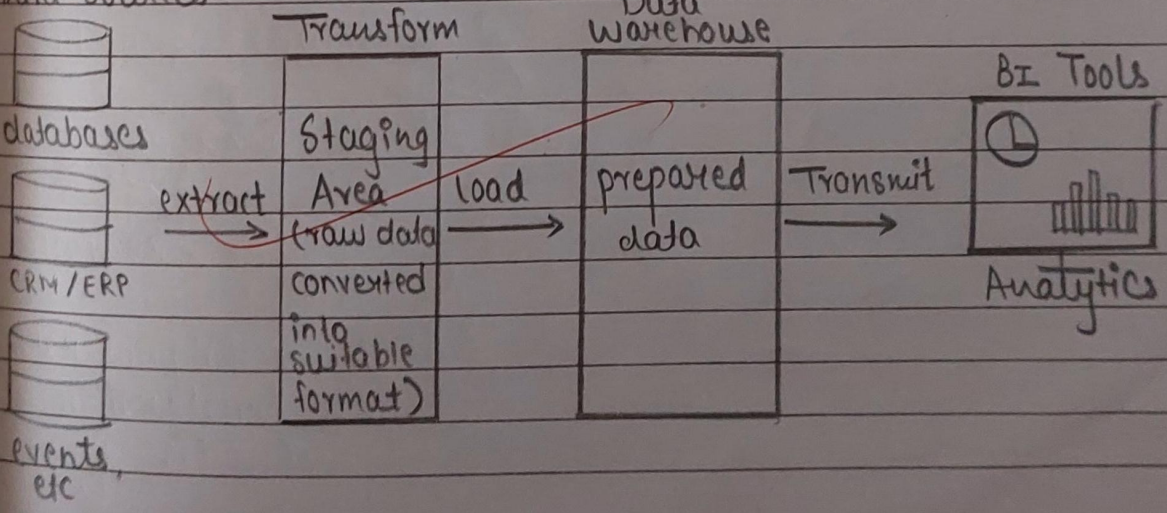
Page No.
Date

* Theory

Extraction gathers data from various sources like files, databases, or semi-structured repositories, validating it, & transferring valid data to a staging area.

Transformation applies predefined rules & functions to the extracted data, including cleansing, encoding, deriving new values, sorting, joining, aggregating, & validating, leading to transformed data ready for storage.

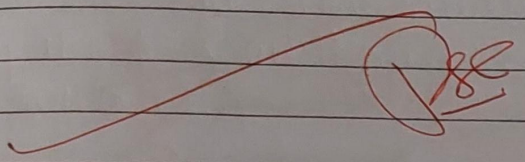
Load places the transformed data into the target destination, often a data mart / warehouse, which may involve appending, refreshing, or overwriting existing data, applying constraints, triggers for data integrity, & executing additional processes like backups or replication.



Page No.
Date

* Conclusion

This way data visualization from ETL process is done.



Date

Assignment 3

* Problem Statement

Perform the extraction Transformation & loading (ETL) process to construct the database in the Sql server / Power BI.

* Title

Perform the extraction Transformation & loading process to construct the database in the sql server / Power BI.

* Prerequisite

Python

* Software Requirements

Jupyter

* Hardware Requirements

PIV, 2 GB RAM, 500 GB HDD

* Outcomes

Page No.
Date

After completion of this assignment students are able to understand how to perform the extraction Transformation & loading process to construct the database in the sql server / Power BI.

* Theory

Step 1 : Data Extraction

It is the first step of ETL

(1) Identify data sources
(2) Extract data using appropriate tools
(3) Validate extracted data
(4) Transfer data to staging area

Step 2 : Data Transformation

(1) Cleanse data to ensure quality
(2) Apply predefined business rules
(3) Join data from multiple sources.
(4) Aggregate data at desired levels.
(5) Validate transformed data.

Step 3 : loading

(1) Design database schema
(2) load transformed data into database tables.

(3) Apply constraints & triggers for data integrity.
(4) Schedule refresh for regular updates.
(5) Monitor & optimize ETL process for performance.

* Conclusion

We have implemented ETL process to construct the database in the Sql server / Power BI

Pse

Page No.
Date

Assignment 4

* Problem Statement

Data Analysis & visualization using advanced excel.

* Title

Data Analysis & visualization using advanced excel.

* Prerequisite

Python

* Software Requirements

Jupyter

* Hardware Requirements

PIV, 2 GB RAM, 500 GB HDD

* Outcomes

After completion we are able to perform data analysis & visualization using advanced excel.

* Theory

Page No.
Date

Data Analysis Techniques

(1) Descriptive Statisics functions like mean and median are provided by excel
(2) Pivot Tables summarize & analyze large datasets dynamically.
(3) Data validation ensures accuracy & consistency in data.

Advanced excel functions & formulas

(1) Array formulas allow for simultaneous calculations on multiple values.
(2) lookup & reference functions retrieve data.
(3) financial functions aid in analyzing financial data.

Data Visualization Techniques

(1) Conditional formatting highlights data patterns using visual cues.
(2) Slicers & timelines enhance dashboard interactivity.
(3) Sparklines provide miniature charts for quick data trend analysis.

Dashboard creation

(1) Dashboards consists of charts, tables, & interactive elements.
(2) Effective dashboard design prioritizes clarity & relevance.

* Conclusion

We have implemented data analysis & visualization using advanced excel.

Assignment 5

* Problem Statement

Perform the data classification algorithm using any classification algorithm.

* Title

Perform the data classification algorithm using any classification algorithm.

* Prerequisite

Python

* Software Requirements

~~Power~~ B ~~Iuy~~ Jupyter

* Hardware Requirements

PIV, 2GB RAM, 500 GB HDD

* Learning Objective

After completion of this assignment we will be able to perform data classification algorithm.

Page No.
Date

Theory

Data Classifications

It is typically a supervised learning task, where the algorithm learns from labeled training data. Features are the attributes or variables that describe each data point, while labels are the categories or classes that data points belong to. The data is divided into training & testing tests. The training set is used to train the classification model, while the testing set is used to evaluate its performance. Classification algorithms learn decision boundaries that separate different classes in the feature space. Common evaluation metrics for classification algorithms include accuracy, precision, Roc Curve, F1 score, recall.

Classification Algorithms

① logistic Regression

It is a linear classification algorithm used for binary classification tasks.

② Decision Trees

Decision trees partition the feature space into

regions & make predictions based on majority class in each regions.

③ Raudom forest

It is an ensemble method that combines multiple decision trees to improve classification performance.

④ Support Vector Machines (SVM)

It finds the optimal hyperplane that separates different classes in the feature space with the maximum margin.

Conclusion

We have implemented data classification successfully.