NMIMS Deemed to be University

Centre of Excellence in Analytics & Data Science

BIG DATA ANALYTICS

IOWA Liquor Sales

By Group 1 Section B

Anuj Shah (B009) - 80812200072
Ayush Gupta (B015) - 80812200011
Havisha Mehta (B024) - 80812200211
Sayali Bhambure (B038) - 80812200125
Shraiya Ranka (B041) - 80812200055

Submitted to
Prof. Cyrus Lentin
Submitted on 14h March 2023

Contents

1. Project Overview

We aim to analyse liquor Sales in IOWA district of USA. Our dataset is obtained from Kaggle having Size – 3.47GB and Shape – 12591077 rows, 24 columns.

Columns = {Invoice/Item Number, Date, Store Number, Store Name, Address, City, Zip Code, Store Location, County Number, County, Category, Category Name, Vendor Number, Vendor Name, Item Number, Item Description, Pack, Bottle Volume (ml), State Bottle Cost, State Bottle Retail, Bottles Sold, Sale (Dollars), Volume Sold (Litres), Volume Sold (Gallons).

We have used HDFS, HIVE,PIG and SQOOP for our big data analytics.

2. Learning Objectives

To utilise Big Data Analytics Skills in understanding of various factors influencing Liquor Sales.

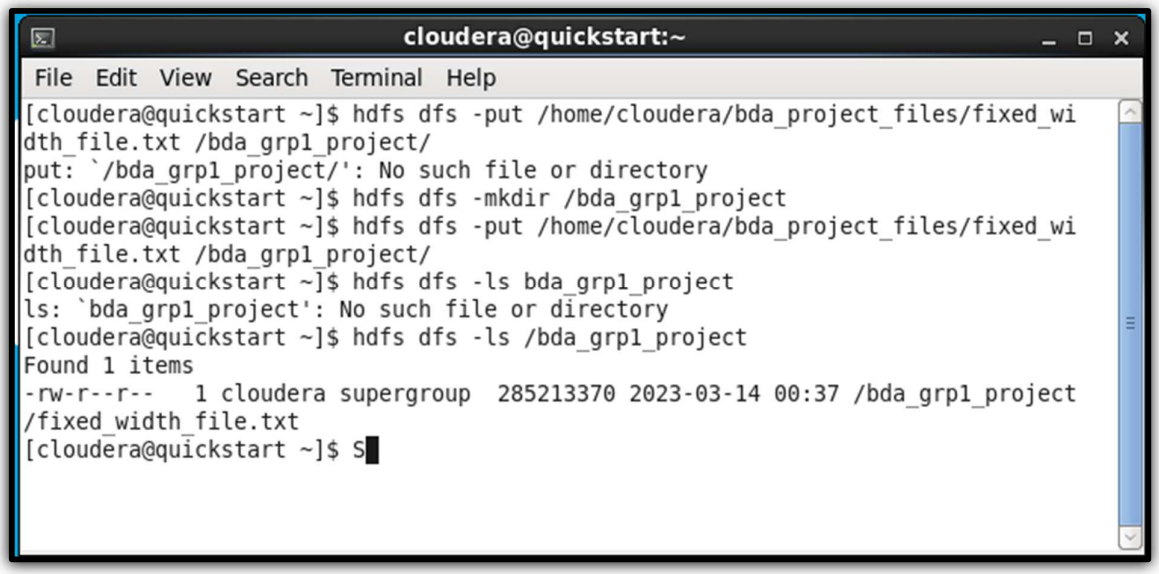
1. Understanding the basics of HDFS and how to store large amounts of data in a distributed file system.
2. Learning how to use Pig to transform and analyze large datasets, and how to write complex Pig scripts to extract relevant information.
3. Understanding how Hive can be used to create tables and manage structured data, and how to write SQL-like queries to analyze the data.
4. Learning how Sqoop can be used to import data from external databases and how to perform data integration tasks with it.
5. Identifying and exploring the key factors that influence liquor sales, such as demographics, geography, seasonality, and marketing campaigns.
6. Applying statistical analysis techniques to identify patterns and trends in the data and gain insights into customer behavior.

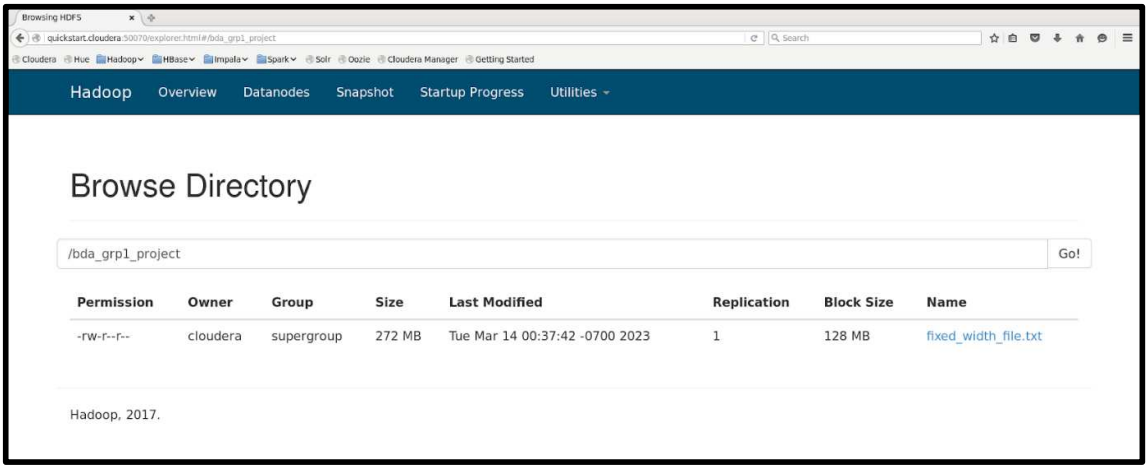
3. Codes and Commands

3.1 HDFS Commands

hdfs dfs -mkdir /bda_grp1_project2

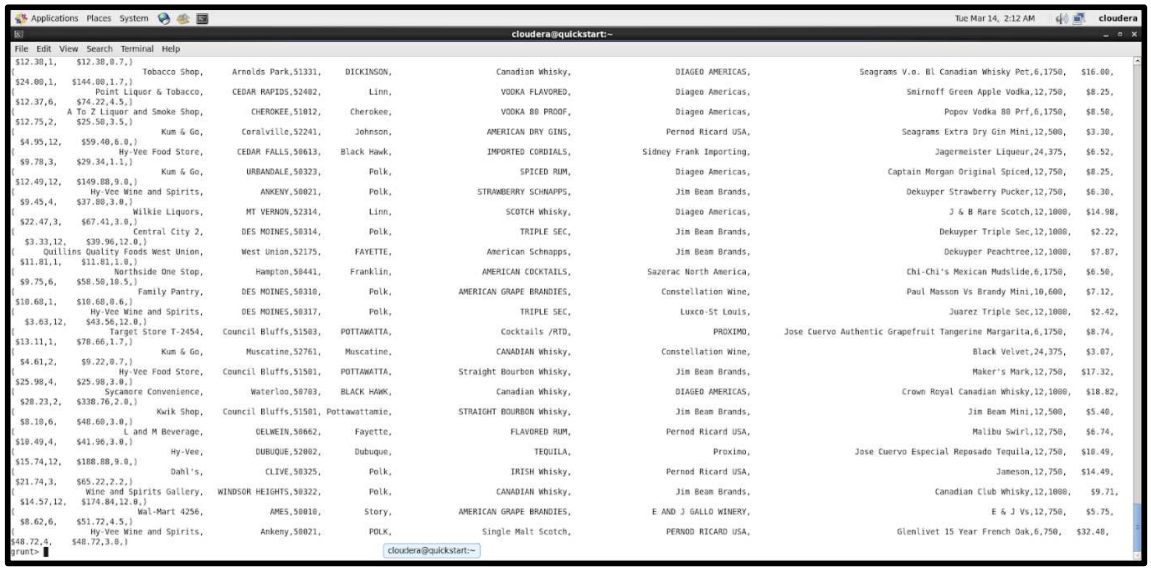
hdfs dfs -put /home/cloudera/bda_project_files2/input/fixed_width_file.txt /bda_grp1_project2/

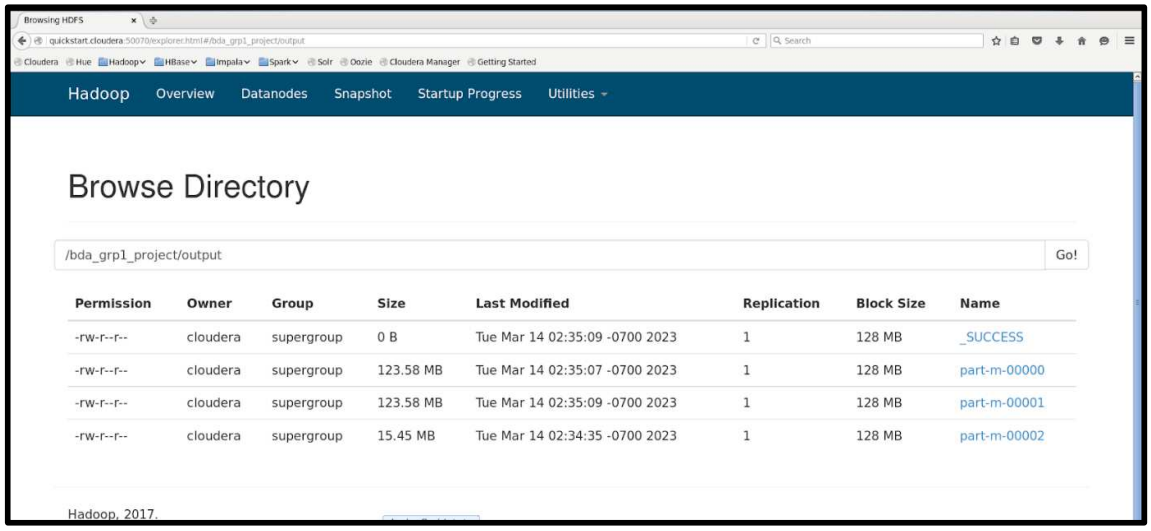




3.2 Pig Commands

```
A = LOAD '/bda_grp1_project2' AS (line:chararray);

B = foreach A GENERATE (chararray)SUBSTRING(line, 0, 39) AS Store_Name,

(chararray)SUBSTRING(line, 39, 57) AS City,

(int)SUBSTRING(line, 57, 63) AS Zip,

(chararray) SUBSTRING(line, 63, 77) AS County,

(chararray) SUBSTRING(line, 77, 114) AS Category_Name,

(chararray) SUBSTRING(line, 114, 151) AS Vendor_Name,

(chararray) SUBSTRING(line, 151, 210) AS Item_Description,

(int) SUBSTRING(line, 210, 214) AS Pack,

(int) SUBSTRING(line, 214, 221) AS Bottle_Volume,

(chararray) SUBSTRING(line, 221, 230) AS State_Bottle_Cost,

(chararray) SUBSTRING(line, 230, 240) AS State_Bottle_Retail,

(int) SUBSTRING(line, 240, 245) AS Bottles_Sold,

(chararray) SUBSTRING(line, 245, 255) AS Sales,

(float) SUBSTRING(line, 255, 262) AS Volume_Sold_Liters,

(float) SUBSTRING(line, 262, 269) AS Volume_Sold_Gallons;

result = FOREACH B GENERATE Store_Name, City, Zip, County, Category_Name, Vendor_Name,
    Item_Description, Pack, Bottle_Volume, State_Bottle_Cost, State_Bottle_Retail, Bottles_Sold,
    Sales, Volume_Sold_Liters, Volume_Sold_Gallons;

DUMP result;

    store result into '/bda_grp1_project2/output' using PigStorage(',');
```


Applications Places System
Tue Mar 14, 2:12 AM
cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
$12.38,1, $12.38,0.7,)
Tobacco Shop, Arnolds Park,51331, DICKINSON, Canadian Whisky, DIAGEO AMERICAS, Seagrams V.o. Bl Canadian Whisky Pet,6,1750, $16.00,
$24.00,1, $144.00,1.7,)
Point Liquor & Tobacco, CEDAR RAPIDS,52402, Linn, VODKA FLAVORED, Diageo Americas, Smirnoff Green Apple Vodka,12,750, $8.25,
$12.37,6, $74.22,4.5,)
A To Z Liquor and Smoke Shop, CHEROKEE,51012, Cherokee, VODKA 80 PROOF, Diageo Americas, Popov Vodka 80 Prf,6,1750, $8.50,
$12.75,2, $25.50,3.5,)
Kum & Go, Coralville,52241, Johnson, AMERICAN DRY GINS, Pernod Ricard USA, Seagrams Extra Dry Gin Mini,12,500, $3.30,
$4.95,12, $59.40,6.0,)
Hy-Vee Food Store, CEDAR FALLS,50613, Black Hawk, IMPORTED CORDIALS, Sidney Frank Importing, Jagermeister Liqueur,24,375, $6.52,
$9.78,3, $29.34,1.1,)
Kum & Go, URBANDALE,50323, Polk, SPICED RUM, Diageo Americas, Captain Morgan Original Spiced,12,750, $8.25,
$12.49,12, $149.88,9.0,)
Hy-Vee Wine and Spirits, ANKENY,50021, Polk, STRAWBERRY SCHNAPPS, Jim Beam Brands, Dekuyper Strawberry Pucker,12,750, $6.30,
$9.45,4, $37.80,3.0,)
Wilkie Liquors, MT VERNON,52314, Linn, SCOTCH Whisky, Diageo Americas, J & B Rare Scotch,12,1000, $14.98,
$22.47,3, $67.41,3.0,)
Central City 2, DES MOINES,50314, Polk, TRIPLE SEC, Jim Beam Brands, Dekuyper Triple Sec,12,1000, $2.22,
$3.33,12, $39.96,12.0,)
Quillins Quality Foods West Union, West Union,52175, FAYETTE, American Schnapps, Jim Beam Brands, Dekuyper Peachtree,12,1000, $7.87,
$11.81,1, $11.81,1.0,)
Northside One Stop, Hampton,50441, Franklin, AMERICAN COCKTAILS, Sazerac North America, Chi-Chi's Mexican Mudslide,6,1750, $6.50,
$9.75,6, $58.50,10.5,)
Family Pantry, DES MOINES,50310, Polk, AMERICAN GRAPE BRANDIES, Constellation Wine, Paul Masson Vs Brandy Mini,10,600, $7.12,
$10.68,1, $10.68,0.6,)
Hy-Vee Wine and Spirits, DES MOINES,50317, Polk, TRIPLE SEC, Luxco-St Louis, Juarez Triple Sec,12,1000, $2.42,
$3.63,12, $43.56,12.0,)
Target Store T-2454, Council Bluffs,51503, POTTAWATTA, Cocktails /RTD, PROXIMO, Jose Cuervo Authentic Grapefruit Tangerine Margarita,6,1750, $8.74,
$13.11,1, $78.66,1.7,)
Kum & Go, Muscatine,52761, Muscatine, CANADIAN Whisky, Constellation Wine, Black Velvet,24,375, $3.07,
$4.61,2, $9.22,0.7,)
Hy-Vee Food Store, Council Bluffs,51501, POTTAWATTA, Straight Bourbon Whisky, Jim Beam Brands, Maker's Mark,12,750, $17.32,
$25.98,4, $25.98,3.0,)
Sycamore Convenience, Waterloo,50703, BLACK HAWK, Canadian Whisky, DIAGEO AMERICAS, Crown Royal Canadian Whisky,12,1000, $18.82,
$28.23,2, $338.76,2.0,)
Kwik Shop, Council Bluffs,51501, Pottawattamie, STRAIGHT BOURBON Whisky, Jim Beam Brands, Jim Beam Mini,12,500, $5.40,
$8.10,6, $48.60,3.0,)
L and M Beverage, DELWEIN,50662, Fayette, FLAVORED RUM, Pernod Ricard USA, Malibu Swirl,12,750, $6.74,
$10.49,4, $41.96,3.0,)
Hy-Vee, DUBUQUE,52002, Dubuque, TEQUILA, Proximo, Jose Cuervo Especial Reposado Tequila,12,750, $10.49,
$15.74,12, $188.88,9.0,)
Dahl's, CLIVE,50325, Polk, IRISH Whisky, Pernod Ricard USA, Jameson,12,750, $14.49,
$21.74,3, $65.22,2.2,)
Wine and Spirits Gallery, WINDSOR HEIGHTS,50322, Polk, CANADIAN Whisky, Jim Beam Brands, Canadian Club Whisky,12,1000, $9.71,
$14.57,12, $174.84,12.0,)
Wal-Mart 4256, AMES,50010, Story, AMERICAN GRAPE BRANDIES, E AND J GALLO WINERY, E & J Vs,12,750, $5.75,
$8.62,6, $51.72,4.5,)
Hy-Vee Wine and Spirits, Ankeny,50021, POLK, Single Malt Scotch, PERNOD RICARD USA, Glenlivet 15 Year French Oak,6,750, $32.48,
$48.72,4, $48.72,3.0,)
grunt>
cloudera@quickstart:~


Browsing HDFS
quickstart.cloudera:50070/explorer.html#/bda_grp1_project/output
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Hadoop Overview Datanodes Snapshot Startup Progress Utilities
Browse Directory
/bda_grp1_project/output
Go!
Permission Owner Group Size Last Modified Replication Block Size Name
-rw-r--r-- cloudera supergroup 0 B Tue Mar 14 02:35:09 -0700 2023 1 128 MB _SUCCESS
-rw-r--r-- cloudera supergroup 123.58 MB Tue Mar 14 02:35:07 -0700 2023 1 128 MB part-m-00000
-rw-r--r-- cloudera supergroup 123.58 MB Tue Mar 14 02:35:09 -0700 2023 1 128 MB part-m-00001
-rw-r--r-- cloudera supergroup 15.45 MB Tue Mar 14 02:34:35 -0700 2023 1 128 MB part-m-00002
Hadoop, 2017.

3.3 Hive Commands

create database liquor;

use liquor;

create table liquor_data(Store_Name string, City string, Zip string, County string, Category_Name string, Vendor_Name string, Item_Description string, Pack int, Bottle_Volume int, State_Bottle_Cost string, State_Bottle_Retail string, Bottles_Sold int, Sales string, Volume_Sold_Liters string, Volume_Sold_Gallons string) row format delimited fields terminated by ',' lines terminated by '\n';

load data inpath '/bda_grp1_project2/output' overwrite into table liquor_data;

1. **Which brands have highest bottle sales and in which counties?**

➔ SELECT county, item_description AS brand, SUM (bottles_sold)*10000 AS TotalBottlesSold

FROM liquor_data

GROUP BY county, Item_Description

ORDER BY TotalBottlesSold DESC

LIMIT 10;

county	brand	totalbottlessold
Polk	Black Velvet	670010000
Polk	Hawkeye Vodka	638110000
Polk	Fireball Cinnamon Whiskey	465510000
Polk	Captain Morgan Spiced Rum	402280000
Polk	Phillips Vodka	375580000
Linn	Hawkeye Vodka	327670000
Linn	Black Velvet	326460000
Polk	Smirnoff Vodka 80 Prf	312770000
Polk	Mccormick Vodka	300470000
Black Hawk	Black Velvet	287750000

10 rows selected (57.256 seconds)

2. **Finding out city wise liquor consumption.**

➔ SELECT city, county, SUM(Volume_Sold_Liters) AS TotalVolume

FROM liquor_data

GROUP BY city, county

LIMIT 20;

city	county	totalvolume
NaN	NaN	328.69999999999953
ALTA	Buena Vista	157.5
Alta	Buena Vista	56.3
COLO	Story	1459.8
Colo	STORY	39.50000000000001
Colo	Story	133.5
DOWS	Wright	552.0
Doon	LYON	14.899999999999999
Dows	WRIGHT	20.699999999999996
Dows	Wright	57.0
ELMA	Howard	2296.4
Elma	HOWARD	129.50000000000003
Elma	Howard	198.5
LEON	Decatur	2574.0000000000005
Leon	DECATUR	295.7999999999996
Leon	Decatur	106.10000000000001
OTHO	Webster	465.9
Otho	WEBSTER	17.099999999999998
Otho	Webster	85.5
Tama	TAMA	16.2

20 rows selected (28.161 seconds)

3. Finding out Brand wise liquor margins (Sale($)/Volume)

➔ SELECT Item_Description AS Brand, (state_bottle_retail - state_bottle_cost) AS Margin

FROM liquor_data

GROUP BY Item_Description

ORDER BY Margin DESC;

brand	margin
Crown Royal Canadian Whisky	86.4
Crown Royal	57.0
Crown Royal Canadian Whisky	48.0
Jameson	45.0
Baileys Original Irish Cream	45.0
Black Velvet	38.7
Black Velvet	38.400000000000006
Barton Vodka	38.400000000000006
Black Velvet	36.75
Cedar Ridge Single Malt Whiskey	36.0

10 rows selected (51.939 seconds)

4. **County-wise liquor consumption in volume**

➔ SELECT County, SUM(Volume_Sold_Liters) AS TotalVolume

FROM liquor_data

GROUP BY County

ORDER BY TotalVolume DESC

LIMIT 10;

county	totalvolume
Polk	1495501.099999466
Linn	649409.6999999792
Scott	495026.9000000331
Black Hawk	420859.6000000223
Johnson	406355.8000000339
Pottawattamie	258820.00000001414
Woodbury	237289.0000000127
Dubuque	231637.6000000125
Story	229712.30000001416
Cerro Gordo	166399.7000000018

5. **What is the average revenue per store with respect to city and county?**

SELECT County, City, AVG(Sales / Bottles_Sold) AS AvgRevenuePerStore

FROM liquor_data

GROUP BY County, City

ORDER BY County, City;

county	city	avgrevenueperstore
Polk	DES MOINES	3801600.0000000005
Polk	WINDSOR HEIGHTS	2979900.0000000005
Polk	WINDSOR HEIGHTS	2956800.0000000005
Crawford	DENISON	2829750.0
Polk	WINDSOR HEIGHTS	2714250.0
Polk	ALTOONA	2310000.0
Shelby	HARLAN	2310000.0
Linn	MARION	2310000.0
Johnson	IOWA CITY	2310000.0
Jones	ANAMOSA	2310000.0

3.4 SQOOP

sqoop export --connect jdbc:mysql://localhost/liquor --username root -P --table new_liquor_data --export-dir /bda_grp1_project2/output --input-fields-terminated-by ',' --lines-terminated-by '\n'

1. **Which item is sold in large packs citywise?**

➔ SELECT City, Item_Description, Pack

FROM liquor_data

WHERE Pack > 24

ORDER BY Pack DESC

LIMIT 10;

2. **What item is available in different variants of bottle volumes?**

➔ SELECT Item_Description, GROUP_CONCAT(DISTINCT Bottle_Volume) AS AvailableVolumes

FROM liquor_data

GROUP BY Item_Description;

4. Summary

Q1. Explain how you used Hadoop for Big Data Analytics.

➔ We downloaded our dataset and imported the csv file in Hadoop, stored it in the directory and using hdfs pig and hive commands we explored our dataset and tried to obtain insights about the dataset.

From a large dataset of almost 12591077 rows and 24 columns we could draw insights on liquor sales in IoWA district, highest volume of liquor is sold in Polk, which means that companies like Diageo can target this county for promoting sales of the liquor being most sold like whiskey.

If we do further deep research taking customers buying these liquors into account and dates on which they purchased liquor we could be able to tell the type of drinks preferred in that county and if there si any seasonality in purchasing drinks.

Demographics of the sales would help in market research and also increase/ decrease in drinking population and further implications on the tax regime on liquor.

Q2. Describe your experience of using Hadoop for analyzing Big Data.

➔ Using Hadoop for Analysing Big Data was a fun learning experience. It was new for us to work on a virtual machine and explore the analytics world. After learning Big Data Analytics we were able to analyse data using virtual machines and it makes the job easier.

One of the main advantages of using Hadoop for big data analysis is its ability to handle data at a large scale. By breaking large datasets into smaller chunks and processing them in parallel across multiple nodes, Hadoop can significantly **reduce the time required** to process and analyze large volumes of data.

Hadoop integrates with a wide range of other data processing and analysis tools, including Hive, Pig, SQL. This makes it a **powerful platform** for performing complex data processing and analysis tasks.

a. **Scalability**: Hadoop is designed to handle large-scale data processing, which makes it ideal for handling big data. It can store and process data across a large number of nodes in a cluster, enabling it to handle data volumes that are too large for traditional data processing systems.
b. **Cost-Effective**: Hadoop is an open-source platform, which means that it is free to use and doesn't require expensive proprietary software licenses. This makes it an affordable option for businesses of all sizes, particularly those that are looking to store and process large volumes of data.
c. **Fault Tolerance**: Hadoop is built to be fault-tolerant, which means that it can continue to function even if one or more nodes in the cluster fail. This helps to ensure the reliability and availability of data processing and analysis, even in the event of hardware failures or other issues.

d. **Flexibility:** Hadoop is designed to work with a wide range of data types and formats, including structured, semi-structured, and unstructured data. This makes it a versatile tool that can be used for a wide range of applications, from data warehousing to machine learning.
e. **Integration with other tools**: Hadoop integrates with a wide range of other data processing and analysis tools, including Hive, Pig, etc. This makes it a powerful platform for performing complex data processing and analysis tasks.