Pandas IO Basics - Tools

Ethans
Learn from experts

Reader functions

- *read_csv*
- *read_excel*
- *read_hdf*
- *read_sql*
- *read_json*
- *read_msgpack* (experimental)
- *read_html*
- *read_gbq* (experimental)
- *read_stata*
- *read_sas*
- *read_clipboard*
- *read_pickle*

Writer functions

- *to_csv*
- *to_excel*
- *to_hdf*
- *to_sql*
- *to_json*
- *to_msgpack* (experimental)
- *to_html*
- *to_gbq* (experimental)
- *to_stata*
- *to_clipboard*
- *to_pickle*

Slide 321

www.ethans.co.in

Read csv

Ethans
Learn from experts

```
df = pd.read_csv(r'C:\ethans\Training\Python\India_Population.csv')
```

```
>>> df.head()
         Date     Value
0  2020-12-31  1380.007
1  2019-12-31  1362.087
2  2018-12-31  1344.401
3  2017-12-31  1326.944
4  2016-12-31  1309.713
>>> df.tail()
         Date     Value
36 1984-12-31   747.000
37 1983-12-31   731.000
38 1982-12-31   715.563
39 1981-12-31   699.938
40 1980-12-31   685.688
```

```
>>> df.describe()
             Value
count    41.000000
mean   1026.812854
std     210.235406
min     685.688000
25%     847.438000
50%    1029.188000
75%    1195.063000
max    1380.007000
```

```
>>> len(df)
41
>>> df.columns
Index([u'Date', u'Value'], dtype='object')
>>> df.set_index('Date', inplace=True)
```

```
>>> df.head()
              Value
Date
2020-12-31  1380.007
2019-12-31  1362.087
2018-12-31  1344.401
2017-12-31  1326.944
2016-12-31  1309.713
```

Slide 322

www.ethans.co.in

Write html

Ethans
Learn From experts

```
>>> df.columns = ['Population']
>>> df.head()
            Population
Date
2020-12-31     1380.007
2019-12-31     1362.087
2018-12-31     1344.401
2017-12-31     1326.944
2016-12-31     1309.713
>>> df.to_html('IndiaPopulation.html')
```

```
>>> df.iloc[1]
Value    1362.087
Name: 2019-12-31 00:00:00, dtype: float64
>>> df[df.Value > 1000]
```

```
>>> df[df.Value > 1000].tail(1)
                Value
Date
1999-12-31  1010.188
```

Date	Population
2020-12-31	1380.007
2019-12-31	1362.087
2018-12-31	1344.401
2017-12-31	1326.944
2016-12-31	1309.713
2015-12-31	1292.707
2014-12-31	1275.921
2013-12-31	1259.353
2012-12-31	1243.000
2011-12-31	1217.438

Concatenating

Ethans
Learn From experts

```
import pandas as pd

df1 = pd.DataFrame({'FSI':[80,85,88,85],
                    'Interest_rate':[2, 3, 2, 2],
                    'PSR':[500, 550, 650, 650]},
                   index = [2001, 2002, 2003, 2004])

df2 = pd.DataFrame({'FSI':[80,85,88,85],
                    'Interest_rate':[2, 3, 2, 2],
                    'PSR':[709, 750, 802, 890]},
                   index = [2005, 2006, 2007, 2008])

df3 = pd.DataFrame({'FSI':[80,85,88,85],
                    'Interest_rate':[2, 3, 2, 2],
                    'Govt_circle_rate':[450, 520, 570, 590]},
                   index = [2001, 2002, 2003, 2004])
```

```
# Concanating df1 and df2, columns are common
concat = pd.concat([df1,df2])
print concat

# Concanating df1 and df3, columns are common
concat = pd.concat([df1,df3])
print concat

# Concatinating df1, df2 and df3, columns are different
concat = pd.concat([df1,df2,df3])
print(concat)
```

Appending

Ethans
Learn from experts

```
import pandas as pd

df1 = pd.DataFrame({'FSI':[80,85,88,85],
                    'Interest_rate':[2, 3, 2, 2],
                    'PSR':[500, 550, 650, 650]},
                   index = [2001, 2002, 2003, 2004])

df2 = pd.DataFrame({'FSI':[80,85,88,85],
                    'Interest_rate':[2, 3, 2, 2],
                    'PSR':[709, 750, 802, 890]},
                   index = [2005, 2006, 2007, 2008])

df3 = pd.DataFrame({'FSI':[80,85,88,85],
                    'Interest_rate':[2, 3, 2, 2],
                    'Govt_circle_rate':[450, 520, 570, 590]},
                   index = [2001, 2002, 2003, 2004])

# Same columns appending
df4 = df1.append(df2)
print(df4)

# Different columns appending
df5 = df1.append(df3)
print(df5)
```


Merging

Ethans
Learn from experts

```
#-------------------------------------------------------------------
# Merging
raw_data = {
        'subjectID': ['1', '2', '3', '4', '5', '7', '8', '9', '10', '11'],
        'Marks': [51, 15, 15, 61, 16, 14, 15, 1, 61, 16]}
df1 = pd.DataFrame(raw_data)
print 'Subjects and Marks ', '--' * 30
print df1

raw_data = {
        'subjectID': ['1', '2', '3', '4', '5'],
        'firstname': ['Alex', 'Amy', 'Allen', 'Alice', 'Ayoung'],
        'lastname': ['Anderson', 'Ackerman', 'Ali', 'Aoni', 'Atiches']}
df2 = pd.DataFrame(raw_data)
print 'Names and SubjectID ', '--' * 30
print df2

#Join
print 'Inner Join ', '--' * 30
print pd.merge(df1, df2, on='subjectID')

#right join
print 'Right Join', '--' * 30
print pd.merge(df1, df2, on='subjectID', how='right')

#Left join
print 'Left Join', '--' * 30
print pd.merge(df1, df2, on='subjectID', how='left')
```

Data Analysis – On movie Data

Ethans
Learn From experts

MovieLens

GroupLens Research has collected and made available rating data sets from the MovieLens web site (http://movielens.org). The data sets were collected over various periods of time, depending on the size of the set. Before using these data sets, please review their README files for the usage licenses and other details.

Help our research lab Please take a short survey about the MovieLens datasets

MovieLens 100K Dataset

Stable benchmark dataset. 100,000 ratings from 1000 users on 1700 movies. Released 4/1998.

- README.txt
- ml-100k.zip (size: 5 MB, checksum)
- Index of unzipped files

Permalink: http://grouplens.org/datasets/movielens/100k/

Analyzing Data Set – User Data

Ethans
Learn From experts

User Data:

'UserId', 'Age', 'Sex', 'Occ', 'Zip'

```
1|24|M|technician|85711
2|53|F|other|94043
3|23|M|writer|32067
4|24|M|technician|43537
5|33|F|other|15213
6|42|M|executive|98101
7|57|M|administrator|91344
8|36|M|administrator|05201
9|29|M|student|01002
10|53|M|lawyer|90703
11|39|F|other|30329
12|28|F|other|06405
```

```
user_col = ['UserId', 'Age', 'Sex', 'Occ', 'Zip']
users = pd.read_csv('u.user', sep='|', names = user_col)
```

Data Set – Rating Data

Ethans
Learn from experts

Rating Data:

'UserId', 'MovieId', 'rating', 'timeStamp'

```
196 242 3   881250949
186 302 3   891717742
22  377 1   878887116
244 51  2   880606923
166 346 1   886397596
298 474 4   884182806
115 265 2   881171488
253 465 5   891628467
305 451 3   886324817
```

```
data_col = ['UserId', 'MovieId', 'rating', 'time']
ratingData = pd.read_csv('u.data', sep='\t', names = data_col)
```


Data Set – Movie Data

Ethans
Learn from experts

Movie Data:

'MovieId', 'title', 'release', 'videoRelease', 'url'

```
1|Toy Story (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Toy%20Story%20
2|GoldenEye (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?GoldenEye%20(1
3|Four Rooms (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Four%20Rooms%
4|Get Shorty (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Get%20Shorty%
5|Copycat (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Copycat%20(1995)
6|Shanghai Triad (Yao a yao yao dao waipo qiao) (1995)|01-Jan-1995||http://us.i
```

```
movie_col = ['MovieId', 'title', 'release', 'videoRelease', 'url']
MovieData = pd.read_csv('u.item', sep='|', names = movie_col, usecols = range(5))
```

Problem Statements and Solutions

Ethans
Learn From experts

1 - Find the 5 top rated movies in the list.

```
movie_rating = pd.merge(MovieData, ratingData)
data = pd.merge(movie_rating, users)

print data.head(5)

most_rated = data.groupby('title').size().sort_values(ascending=False)[:5]
print most_rated

most_rated.plot()
show()
```

2 – Which age group users provide the maximum ratings?

```
####------------------------------------------------------------
###2
labels = ['0-9', '10-19', '20-29', '30-39', '40-49', '50-59', '60-69', '70-79']
data['age_group'] = pd.cut(data.Age, range(0, 81, 10), right = False, labels = labels)
print data.head(5)
print data.groupby('age_group').size().sort_values(ascending=False)[:1]
```


pandas_datareader

Ethans
Learn From experts

```
__author__ = "Ethan's"

import pandas as pd
import datetime
from pandas_datareader import data
import matplotlib.pyplot as pyplot

startDate = datetime.datetime(2015, 1, 1)
endDate = datetime.datetime(2016, 1, 1)
df = data.DataReader("AAPL", "google", startDate, endDate)

print(df.head())
df['High'].plot()
pyplot.legend()
pyplot.show()
```

Objective – Module 13

Ethans
Learn from experts

Hadoop data Processing with Python

- Introduction of Big Data
- Why Big Data
- Hadoop Eco system
- Understanding problem statements
- Market data Analysis with Python
- HDFS file system
- Cloudera Cluster of single node
- Map Reduce using Python
- Introduction to MrJob Package

Slide 333

What is Big Data?

Ethans
Learn from experts

Dictionary says: Extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions.

Some Examples:

- Air Bus A380 engine generates apx 10 TB of data every 30 Min.
- Facebook generates apx 20 TB of data per day .
- Twitter generates apx 20 TB of data per day.
- Google processes apx 20 PB a day (2008)
- New York Stock Exchange apx 1TB of data everyday.

Slide 334

We are Living in data driven world

Ethans
Learn From experts

Telecom
Huge data all details records, messaging, whatapp, Hike and so on..

- Science
  Huge Data from environmental data, transportation data, sensors, CCTV and so on
- Humanities and Social Sciences
  Scanned books, Gmail, Facebook, twitter, social interactions data, new technology like GPS>
- Business & Commerce
  Sales, Online advertising, stock market transactions, airline traffic etc
- Entertainment
  Internet images, Movies, MP3 files etc
- Medicine
  MRI & CT scans, patient records, DNA profiles, step cells

Slide 335

www.ethans.co.in

Why Big Data is important?

Ethans
Learn From experts

The importance of big data doesn't revolve around how much data you have, but what you do with it. You can take data from any source and analyze it to find answers that enable:
1) cost reductions, 2) time reductions, 3) new product development and optimized offerings, and 4) smart decision making.

When you combine big data with high-powered analytics, you can accomplish business-related tasks such as:

- Determining root causes of failures, issues and defects in near-real time – Examples of Telecom and Airline
- Generating coupons at the point of sale based on the customer's buying habits. Examples of Online advertising
- Recalculating entire risk portfolios in minutes. – Examples of Share Market
- Detecting fraudulent behaviour before it affects your organization. – Examples of Anti Virus software's and many

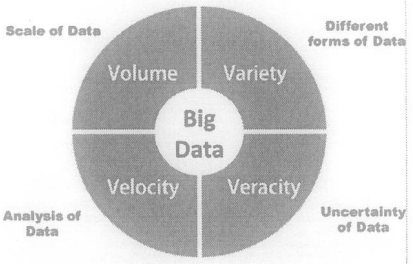
Slide 336

www.ethans.co.in

Big Data Vectors

Ethans
Learn from experts

- Volume - To many bytes
- Velocity - To high rate
- Variety - To many sources
- Veracity – Uncertainty of data

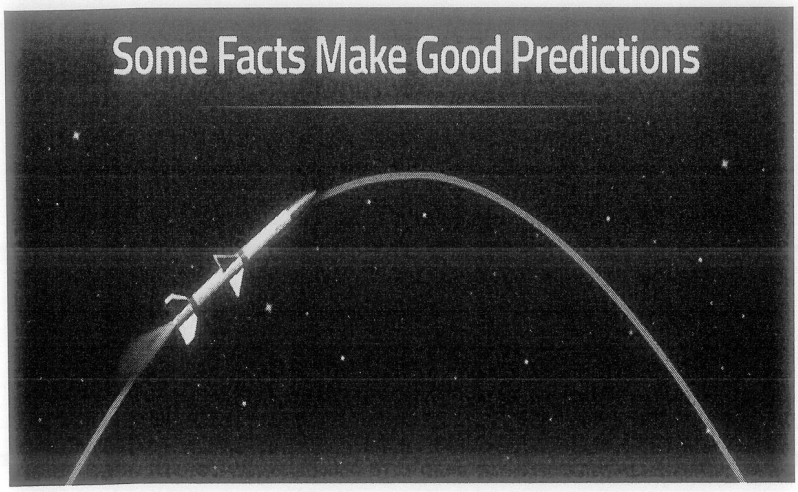


Slide 337

www.ethans.co.in

Let's have some predictions?

Ethans
Learn from experts

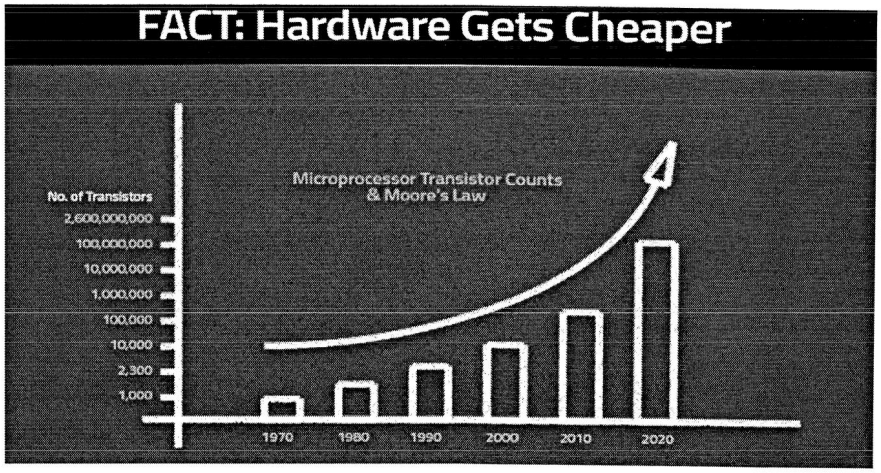


Slide 338

www.ethans.co.in

Moore's Law – What is future?

Ethans
Learn from experts



Slide 339

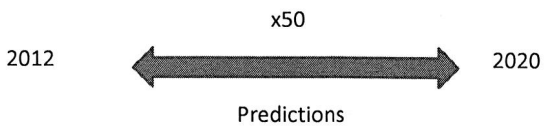
www.ethans.co.in

Predictions

Ethans
Learn from experts

- As of 2009, the entire World Wide Web was estimated to contain close to 500 Exabyte's. This is a half Zeta byte

- The total amount of global data is grown to 2.7 Zeta bytes during 2012. This is 48% up from 2011



Slide 340

www.ethans.co.in

Where we can store data?

Ethans
Learn from experts

RDBMS

- High Concurrency
- TB Storage
- Indexed reads
- Efficient updates

Analytic Appliances

- Scalable
- Medium Concurrency
- High Volume Processing (Postgres)

NoSQL

- Highly Scalable
- High Concurrency
- Storage Options
- Updates
- Real-time Capable
- Rudimentary

Hadoop

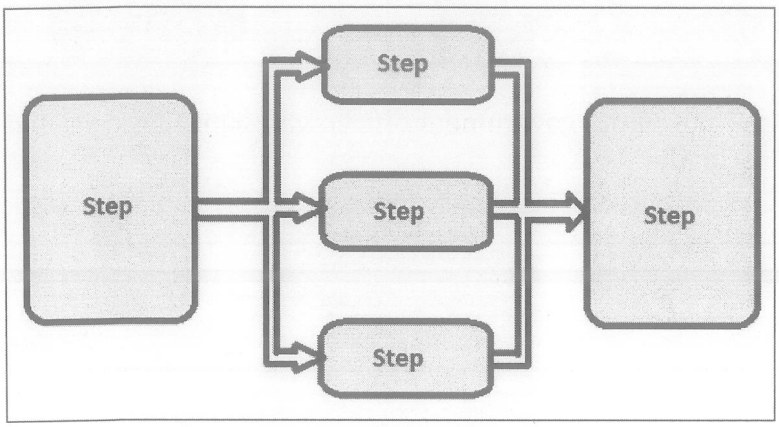
- Highly scalable
- Low concurrency
- Distributed Storage
- Complex Access
- Security (TBD)

Slide 343

www.ethans.co.in

How we process them?

Ethans
Learn from experts



Slide 344

www.ethans.co.in

Challenges

Ethans
Learn From experts

Challenges are huge, Just take an example of internet:

- Per the survey: 250 Crores are using internet and the count will shoot every year.
- Every time you login to Facebook you create a log, if you login 5 times daily : 5 logs
- Similar for whatapp :20 Logs
- Similar for Gmail!
- Similar for Google!
- Similar for many products!

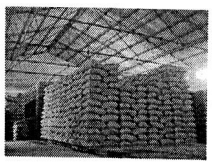
- And Data is valuable !!

Two Challenges we come across: How to store data and How to process Data?

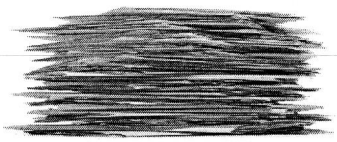
Real world Examples

Ethans
Learn From experts

- Think about a farmer who crop rice on the fields?

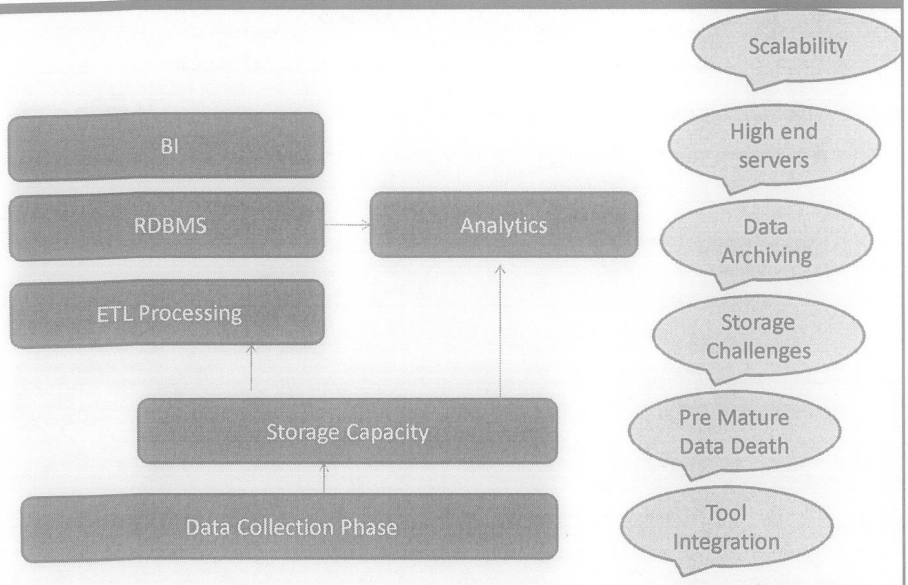


- Think about government officer who signs files daily and regularly?



Challenges with the current infrastructure

Ethans
Learn from experts



Slide 345

www.ethans.co.in

HENCE BIG-DATA

Ethans
Learn from experts

Here is the conclusion.

- We need storage space as much we can!
- We need faster processing!
- We need to process anything!

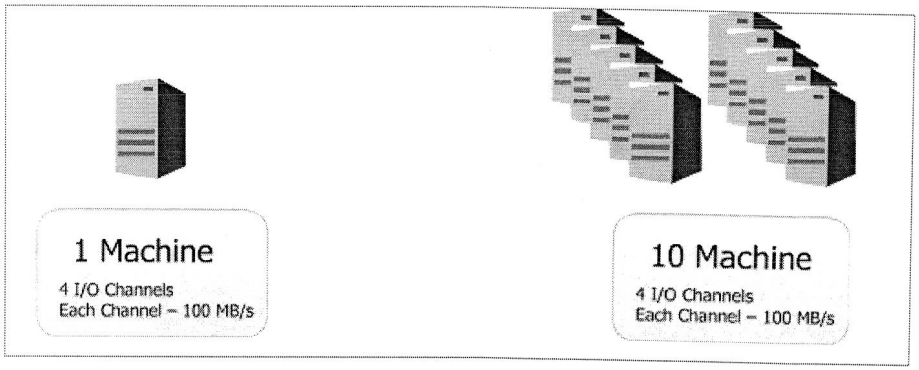
- Volume - To many bytes
- Velocity - To high rate
- Variety - To many sources

Slide 346

www.ethans.co.in

Why DFS?

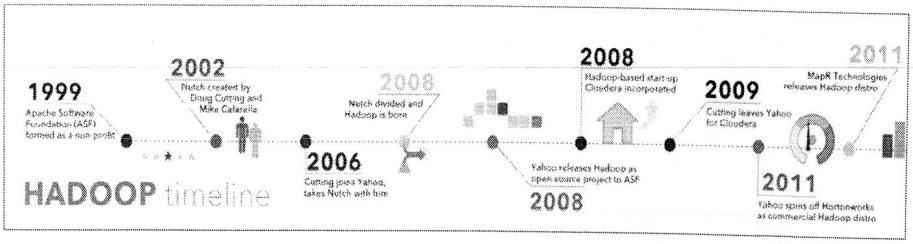
Ethans
Learn From experts



What is Hadoop?

Ethans
Learn From experts

Per SAS - Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs



Why Hadoop? (As per SAS)

Ethans
Learn from experts

Ability to store and process huge amounts of any kind of data, quickly. With data volumes and varieties constantly increasing, especially from social media and the Internet of Things (IoT), that's a key consideration.

Computing power. Hadoop's distributed computing model processes big data fast. The more computing nodes you use, the more processing power you have.

Fault tolerance. Data and application processing are protected against hardware failure. If a node goes down, jobs are automatically redirected to other nodes to make sure the distributed computing does not fail. Multiple copies of all data are stored automatically.

Flexibility. Unlike traditional relational databases, you don't have to preprocess data before storing it. You can store as much data as you want and decide how to use it later. That includes unstructured data like text, images and videos.

Low cost. The open-source framework is free and uses commodity hardware to store large quantities of data.

Scalability. You can easily grow your system to handle more data simply by adding nodes. Little administration is required.

Hadoop Distribution

Ethans
Learn from experts

Cloudera
Cloudera distributes a platform of open-source projects called Cloudera's Distribution including Apache Hadoop or CDH.

Hortonworks
Major contributors to Apache Hadoop and dedicated to working with the community to make Apache Hadoop more robust and easier to install, manage, use, integrate and extend.

IBM InfoSphere BigInsights
brings the power of Apache Hadoop to the enterprise

MapR
sells a high performance map-reduce framework based on Apache Hadoop that includes many of the standard eco-system components.

HDInsight
Managed Apache Hadoop, Spark, Hbase and Storm made easy

Ethans
Learn from experts

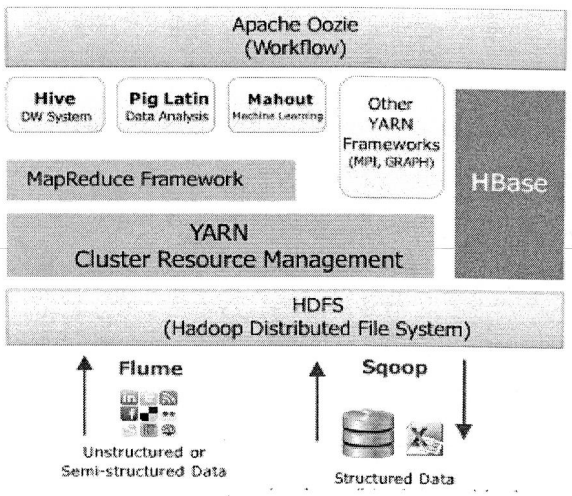
RDBMS		Hadoop
Structured	Data Types	Structure, Semi and Unstructured
Limited	Data Processing	Distributed processing
Standards	Governance	Loosely Standards
Required on write	Schema	Required on Read
Write many read many	Speed	Write once read many
Licence	cost	Open Source
OLTP ACID Transactions	Best Use	Data discovery Heavy Data processing Massive Storage/ Analytics

Slide 351

www.ethans.co.in

Hadoop Eco System

Ethans
Learn from experts



Slide 352

www.ethans.co.in

Ethans
Learn from experts

Slide 353

www.ethans.co.in

Ethans
Learn from experts

Sears

Sears (NASDAQ:SHLD) is an icon of the American business landscape. The company was founded by Richard Warren Sears and Alvah Curtis Roebuck in 1886. At its apogee, Sears was the biggest retailer in the United States. Due to its prosperity, the firm decided to build the largest building in America: the Sears Tower located in Chicago. At completion in 1973, it surpassed the World Trade Center towers and became the tallest building in the world. Now, Sears is not the firm that it once was. It is now the 13th largest retailer in the United States based on annual revenue behind Costco, Wal-Mart and Best buy for example.

Improving customer loyalty, and with it sales and profitability, is desperately important to Sears as it faces fierce competition from Wal-Mart and Target, as well as online retailers such as Amazon.com. While revenue at Sears has declined, from $50 billion in 2008 to $42 billion in 2011, big-box rivals Wal-Mart and Target have grown steadily, and they're far more profitable. Meantime, Amazon has gone from $19 billion in revenue in 2008 to $48 billion last year, passing Sears for the first time.

Slide 354

www.ethans.co.in

Ethans
Learn from experts

Sears → hadoop

Enter Hadoop, an open source data processing platform gaining adoption on the strength of two promises: ultra-high scalability and low cost compared with conventional relational databases. Hadoop systems at 200 terabytes cost about one-third of 200-TB relational platforms, and the differential grows as scale increases into the petabytes, according to Sears. With Hadoop's massively parallel processing power, Sears sees little more than one minute's difference between processing 100 million records and 2 billion records.

Source: http://www.informationweek.com/it-leadership/why-sears-is-going-all-in-on-hadoop/d/d-id/1107038

Slide 355 www.ethans.co.in

Ethans
Learn from experts

US Presidential Elections – Problem Statement:

The 2000 and 2004 presidential elections in the United States were very close.

The largest percentage of the popular vote that any candidate received was 50.7% and the lowest was 47.9%. If a percentage of the voters were to have switched sides, the outcome of the elections would have been different. (Source: Machine learning in action)

There are small groups of voters who, when properly appealed to, will switch sides. These groups may not be huge, but with such close races, they may be big enough to change the outcome of the election.1

How do you find these groups of people, and how do you appeal to them with a limited budget?

Slide 356 www.ethans.co.in

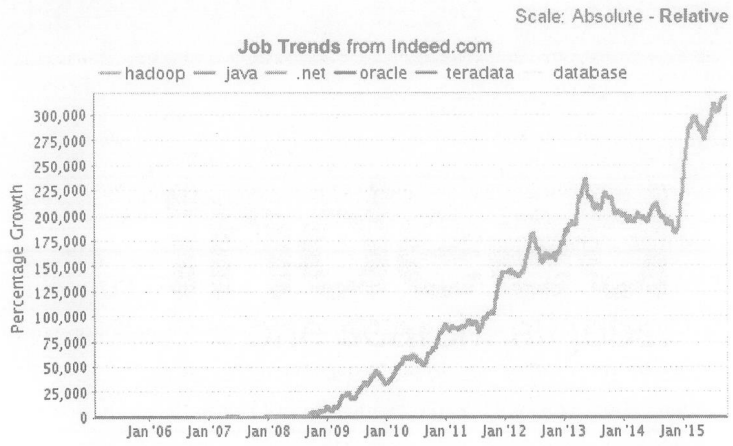
Ethans
Learn from experts

US 2012 Election

- Predictive modeling, clustering modeling and data mining for individual add and do the Analysis on the top of it.
- Hits on mybarackobama.com and identify the target audience.
- Drive traffic to other campaign sites
- Facebook page (33 million "likes")
- YouTube channel (240,000 subscribers and 246 million page views).
- A contest to dine with Sarah Jessica Parker
- Every single night, the team ran 66,000 computer simulations.
- Amazon web services

Ethans
Learn from experts

Job Trend from Indeed



Introduction to HDFS

Ethans
Learn from experts

- Hadoop framework comes with a distributed filesystem called HDFS, which stands for Hadoop Distributed Filesystem.
- HDFS is a Hadoop Flagship filesystem
- It's a file system specially designed for storing very large files with streaming data access patterns, running on clusters of commodity hardware.

Introduction to HDFS

Ethans
Learn from experts

- "Very large size" in this context means files of megabytes, gigabytes , terabytes or peta bytes in size.
- There are many Hadoop clusters store and processing peta bytes of data every day.
- As Moving Computation is Cheaper than Moving Data
- A computation requested by application is much more efficient if it is executed near the data it operates on. When the size of the data set is huge this minimizes network congestion and increases the overall throughput of the system. The assumption is that it is often better to migrate the computation closer to where the data is located rather than moving the data to where the application is running.