Programming for Big Data

Saeed Iqbal Khattak

Centre for Healthcare Modelling & Informatics
Faculty of Information Technology,
University of Central Punjab, Lahore

March 11, 2021

Outline

- Course Outline
- What is Data and Big Data
- Generating Big Data?
- Why Big Data?
- Why Programming for Big Data?
- What is Data Science
- Importance of Data Science

Course Outline

Assessment Instruments with Weights

- Assignments 10%
- Quizzes 10%
- Project 10%
- Midterm 20%
- Class Participation 5%
- Final Term 45%

Recommended Books

1. Mining of Massive Datasets
2. Data Analysis with open source tools
3. Learning Apache Spark with Python
4. Python for Data Analysis
5. Python for Programmers
6. Big Data, Mining, and Analytics

What is Data

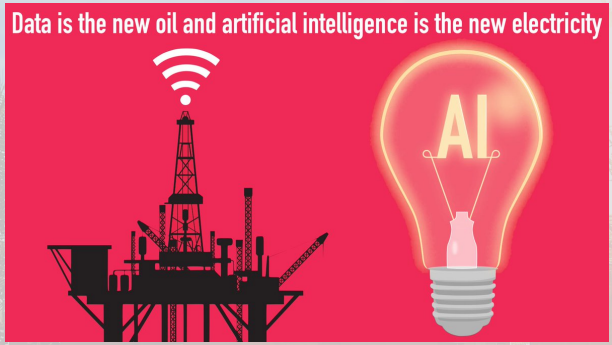


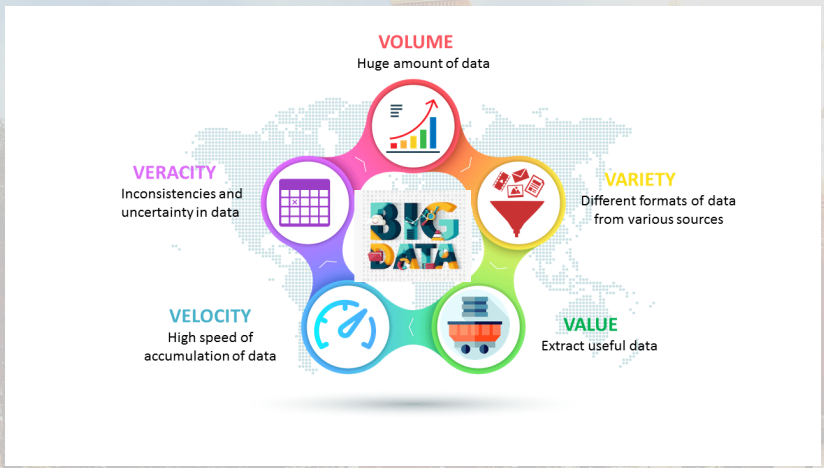
Figure: "Data is Oil" to "AI is Electricity" [1]

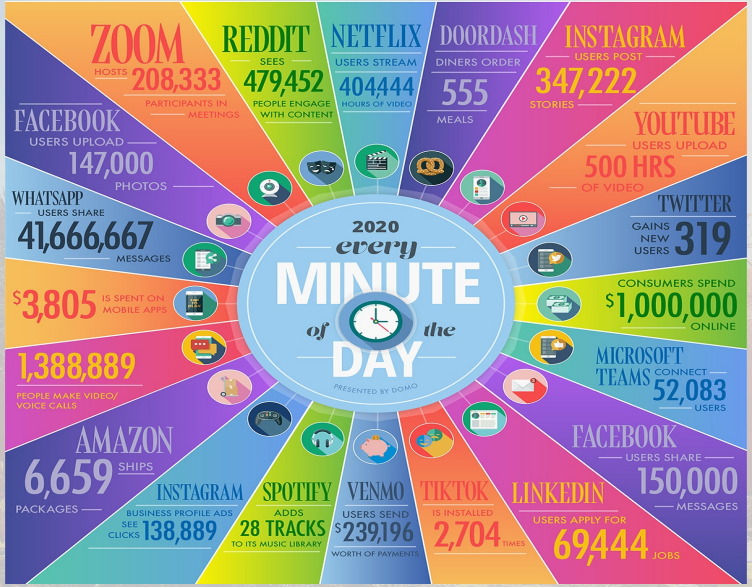
- Clive Humby, UK Mathematician and architect of Tesco's Clubcard, said in 2006.
  - "Data is the new oil" [1].
- Andrew Ng gave the phrase in 2016:
  - "AI is the new Electricity" [1].

What is Big Data

- "Big Data" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it...

V's of Big Data




ZOOM
HOSTS 208,333
PARTICIPANTS IN MEETINGS
REDDIT
SEES 479,452
PEOPLE ENGAGE WITH CONTENT
NETFLIX
USERS STREAM 404,444
HOURS OF VIDEO
DOORDASH
DINERS ORDER 555
MEALS
INSTAGRAM
USERS POST 347,222
STORIES
FACEBOOK
USERS UPLOAD 147,000
PHOTOS
YOUTUBE
USERS UPLOAD 500 HRS
OF VIDEO
WHATSAPP
USERS SHARE 41,666,667
MESSAGES
TWITTER
GAINS NEW USERS 319
2020 every MINUTE of the DAY
PRESENTED BY DOMO
$3,805 IS SPENT ON MOBILE APPS
CONSUMERS SPEND $1,000,000 ONLINE
1,388,889
PEOPLE MAKE VIDEO/ VOICE CALLS
MICROSOFT TEAMS CONNECT 52,083 USERS
AMAZON
6,659 SHIPS
PACKAGES
FACEBOOK
USERS SHARE 150,000
MESSAGES
INSTAGRAM
BUSINESS PROFILE ADS SEE CLICKS 138,889
SPOTIFY
ADDS 28 TRACKS
TO ITS MUSIC LIBRARY
VENMO
USERS SEND $239,196
WORTH OF PAYMENTS
TIKTOK
IS INSTALLED 2,704 TIMES
LINKEDIN
USERS APPLY FOR 69,444 JOBS

Big Data Generation Model & Problem?

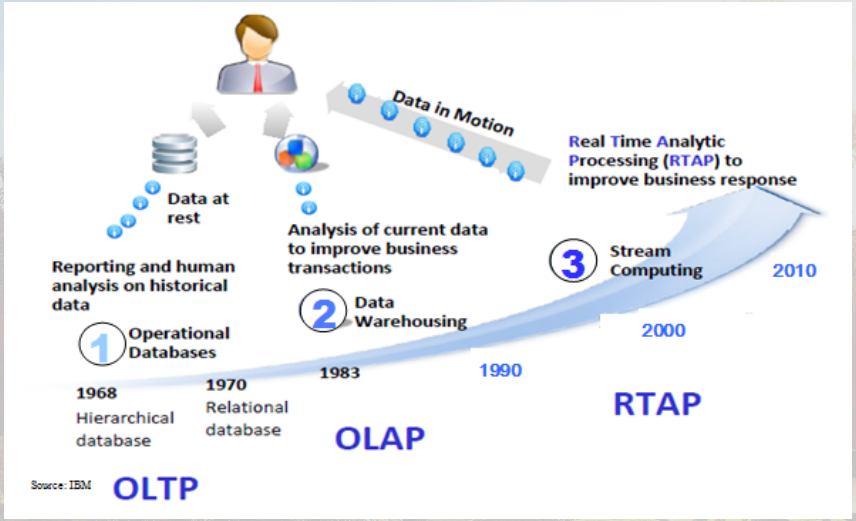
- The Model of Generating/Consuming Data has Changed

Old Model: Companies are generating data and all others are consuming data.

New Model: All of us are generating data and all of us are consuming data.

- The progress and innovation is no longer hindered by the ability to collect data.
- **Problems/Issues??**
  1. Ability to **manage**, **analyze**, **summarize** and **visualize**?
  2. Discover **knowledge** from the collected data in **timely**.

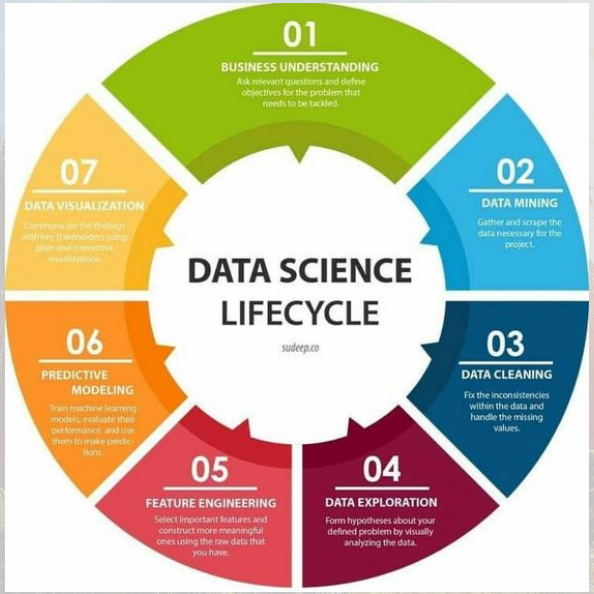
Big Data Harnessing?



Python

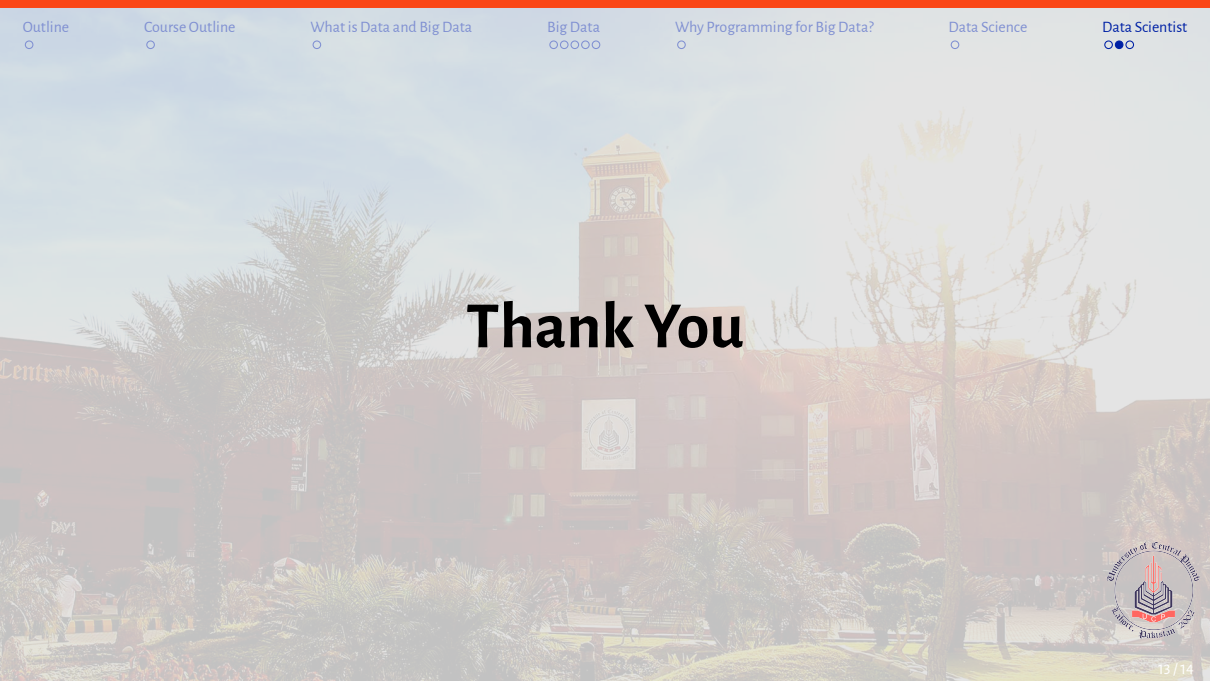
Language Ranking: Trending

Rank	Language	Type	Score
1	Python		100.0
2	Java		94.9
3	C		91.4
4	C++		87.5
5	JavaScript		74.9
6	C#		74.1
7	Go		69.3
8	R		68.8
9	Ruby		67.1
10	Dart		67.1


01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.
DATA SCIENCE LIFECYCLE
sudeep.co

DATA SCIENTIST
SALARY
Average salary of data scientists is
$120,000/year
BENEFITS
• Harvard Business Review called data science the "Sexiest Job of the 21st Century"
• One of the fastest growing careers in the United States
• 94% of data science graduates have found jobs since 2011
RESPONSIBILITIES
• Conduct research
• Extract, clean, and analyze data from varied sources
• Solve problems
• Build automation tools
• Communicate findings to management
EDUCATION
• 88% of all data scientists have at least a Master's degree
• 46% of data scientists have a PhD
SKILLS
• Programming languages (R, Python, SQL, Hive, etc.)
• Statistics
• Multivariable calculus and linear algebra
• Machine learning
• Software engineering
• Wrangle, visualize, and communicate data to management
CAREER POSSIBILITIES
• The majority of data scientists work in the technology industry.
• Other options include marketing, consulting, healthcare and pharmaceuticals, finance,

Thank You



References I

gaurav Tripathi, "Google AI algorithm masters ancient game of Go." https://www.gauravtripathi.com/2018/07/09/data-is-oil-ai-is-electricity-decade-of-evolution/, July 9, 2018. [Online; accessed 27-March-2019].