Big Data Analysis
(MA60306)

Bibhas Adhikari

Spring 2022-23, IIT Kharagpur

Lecture 2
January 5, 2022

Big Data Analysis

Data science in industry (For the course project)

- Setting a research goal - timeline, deliverable

Big Data Analysis

Data science in industry (For the course project)

- ▷ Setting a research goal - timeline, deliverable
- ▷ Retrieving/collecting data - access

Big Data Analysis

Data science in industry (For the course project)

- Setting a research goal - timeline, deliverable
- Retrieving/collecting data - access
- Data preparation: *cleansing*, *integration*, *transformation*

Big Data Analysis

Data science in industry (For the course project)

- ▷ Setting a research goal - timeline, deliverable
- ▷ Retrieving/collecting data - access
- ▷ Data preparation: *cleansing*, *integration*, *transformation*
- ▷ Data exploration: Exploratory Data Analysis (EDA), finding outliers

Big Data Analysis

Data science in industry (For the course project)

- ▷ Setting a research goal - timeline, deliverable
- ▷ Retrieving/collecting data - access
- ▷ Data preparation: *cleansing*, *integration*, *transformation*
- ▷ Data exploration: Exploratory Data Analysis (EDA), finding outliers
- ▷ Data modeling: statistics, machine learning, operations research etc.

Big Data Analysis

Data science in industry (For the course project)

- Setting a research goal - timeline, deliverable
- Retrieving/collecting data - access
- Data preparation: *cleansing, integration*, *transformation*
- Data exploration: Exploratory Data Analysis (EDA), finding outliers
- Data modeling: statistics, machine learning, operations research etc.
- Delivering the observation

About the course project

Follow the above steps and prepare the course project report in Latex

- There can be at most 10 of you in a group

About the course project

Follow the above steps and prepare the course project report in Latex

- There can be at most 10 of you in a group
- The report should be 15 or so pages long and must contain original research (data)

About the course project

Follow the above steps and prepare the course project report in Latex

- There can be at most 10 of you in a group
- The report should be 15 or so pages long and must contain original research (data)
- Use
  documentclass[11pt]article for writing the report

About the course project

Follow the above steps and prepare the course project report in Latex

- There can be at most 10 of you in a group
- The report should be 15 or so pages long and must contain original research (data)
- Use
  documentclass[11pt]article for writing the report
- The proposal is due on Thursday, January 19. It should consist of a title, a paragraph or two on what you plan for the project, and a preliminary list of references

About the course project

Follow the above steps and prepare the course project report in Latex

- There can be at most 10 of you in a group
- The report should be 15 or so pages long and must contain original research (data)
- Use
  documentclass[11pt]article for writing the report
- The proposal is due on Thursday, January 19. It should consist of a title, a paragraph or two on what you plan for the project, and a preliminary list of references
- The report is due on Monday, April 11

About the course project

Follow the above steps and prepare the course project report in Latex

- There can be at most 10 of you in a group
- The report should be 15 or so pages long and must contain original research (data)
- Use
  documentclass[11pt]article for writing the report
- The proposal is due on Thursday, January 19. It should consist of a title, a paragraph or two on what you plan for the project, and a preliminary list of references
- The report is due on Monday, April 11
- Presentations of the project outcomes will be 12-15 minutes long

About the course project

Follow the above steps and prepare the course project report in Latex

- There can be at most 10 of you in a group
- The report should be 15 or so pages long and must contain original research (data)
- Use
  documentclass[11pt]article for writing the report
- The proposal is due on Thursday, January 19. It should consist of a title, a paragraph or two on what you plan for the project, and a preliminary list of references
- The report is due on Monday, April 11
- Presentations of the project outcomes will be 12-15 minutes long
- Topics: as discussed in the class

Big Data Analysis

Big data ecosystem

- Distributed file systems - Hadoop file system (HDFS), *Red Hat Cluster File System*, *Ceph File System*, *Tachyon File System*

Big Data Analysis

Big data ecosystem

- Distributed file systems - Hadoop file system (HDFS), *Red Hat Cluster File System*, *Ceph File System*, *Tachyon File System*
- Distributed programming framework

Big Data Analysis

Big data ecosystem

- Distributed file systems - Hadoop file system (HDFS), *Red Hat Cluster File System*, *Ceph File System*, *Tachyon File System*
- Distributed programming framework
- Data integration framework - *Apache Sqoop*, *Apache Flume excel*

Big Data Analysis

Big data ecosystem

- Distributed file systems - Hadoop file system (HDFS), *Red Hat Cluster File System*, *Ceph File System*, *Tachyon File System*
- Distributed programming framework
- Data integration framework - *Apache Sqoop*, *Apache Flume excel*
- Machine learning frameworks
  - *PyBrain* for neural networks
  - NLTK or Natural Language Toolkit
  - *TensorFlow*

Big Data Analysis

Big data ecosystem

- Distributed file systems - Hadoop file system (HDFS), *Red Hat Cluster File System*, *Ceph File System*, *Tachyon File System*
- Distributed programming framework
- Data integration framework - *Apache Sqoop*, *Apache Flume excel*
- Machine learning frameworks
  - *PyBrain* for neural networks
  - NLTK or Natural Language Toolkit
  - *TensorFlow*
- *No*SQL databases - endless growth of data