DEVELOPING BIG DATA CURRICULUM WITH OPEN SOURCE INFRASTRUCTURE

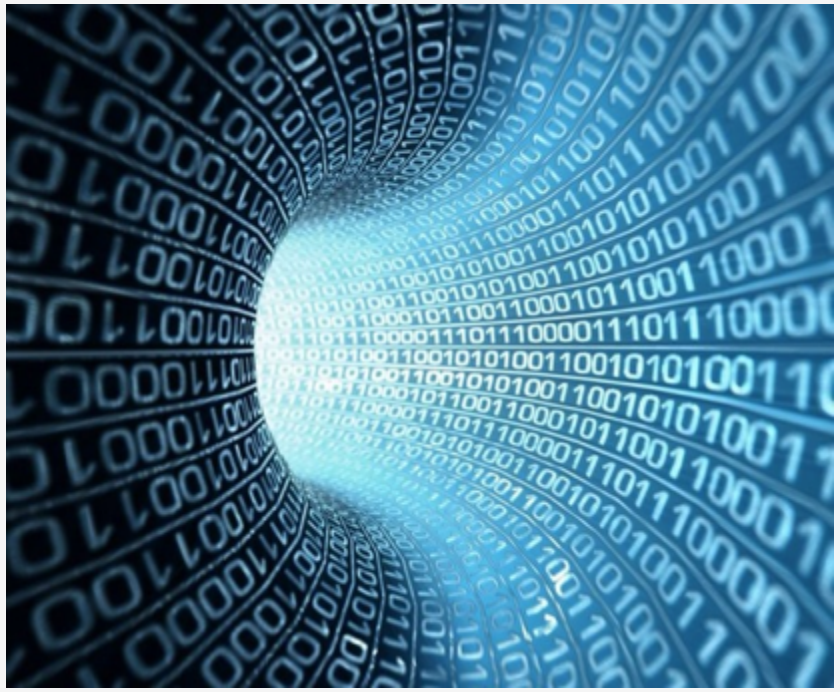
- ANURAG NAGAR
UNIVERSITY OF TEXAS AT DALLAS

UT DALLAS

Big Data Education

- One of the most popular courses at the graduate & undergraduate level.
- Huge demand from industry.
- Curriculum involves study of distributed and parallel computing systems.
- Cutting edge infrastructure is essential to give students rich experience.

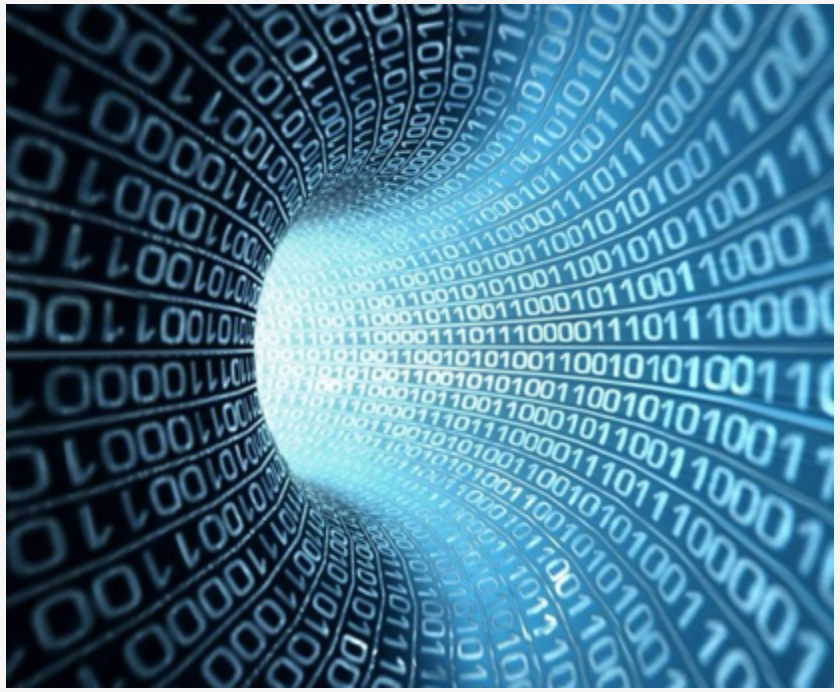
Infrastructure



Infrastructure needed:

- Hadoop Distributed File System
  - Ability to run large MapReduce jobs
- Apache Spark for analytics and machine learning
- Apache Pig for workflow modeling
- Apache Hive for data warehousing
- High availability Cassandra cluster

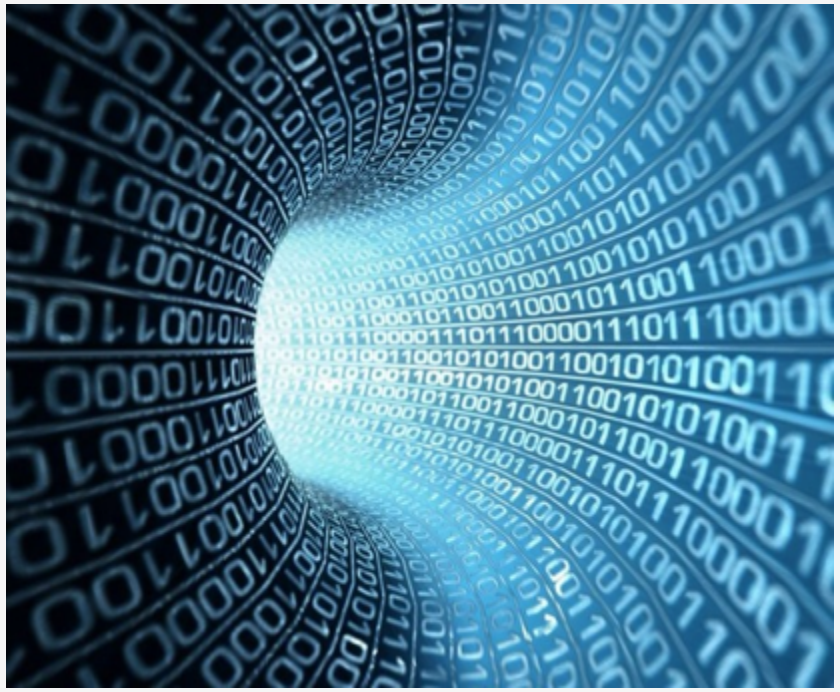
Open Source Solutions



Hadoop:

- Apache Spark install on personal machine
- Cloudera Virtual Machine
- Cloudera Docker
- Amazon EC2 / EMR
  - free credits for students
- Microsoft HDInsight
  - free credits

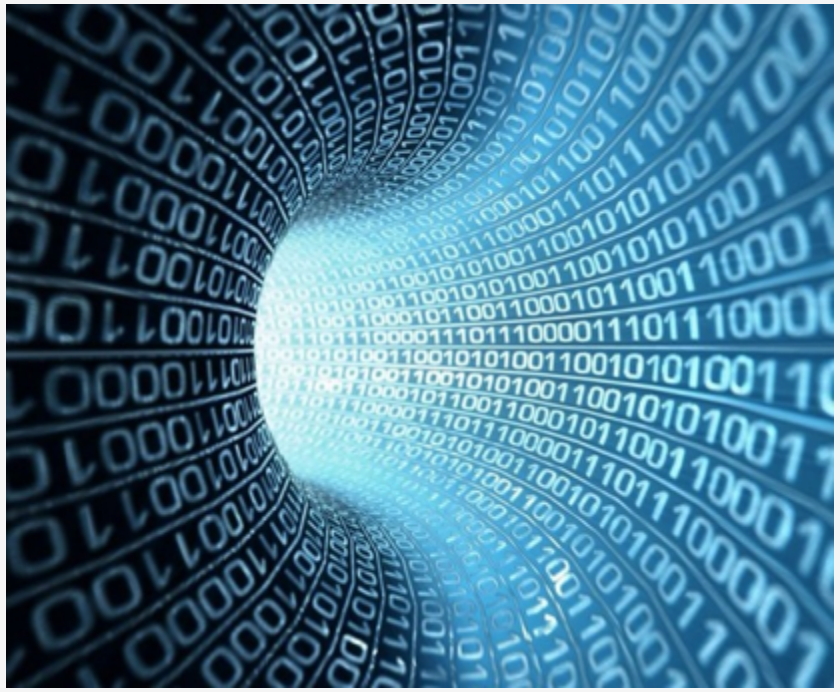
Open Source Solutions



Spark:

- Databricks community edition
  - free 6 GB cluster
- Install Spark on personal machine
- Amazon EC2
  - free credits for students

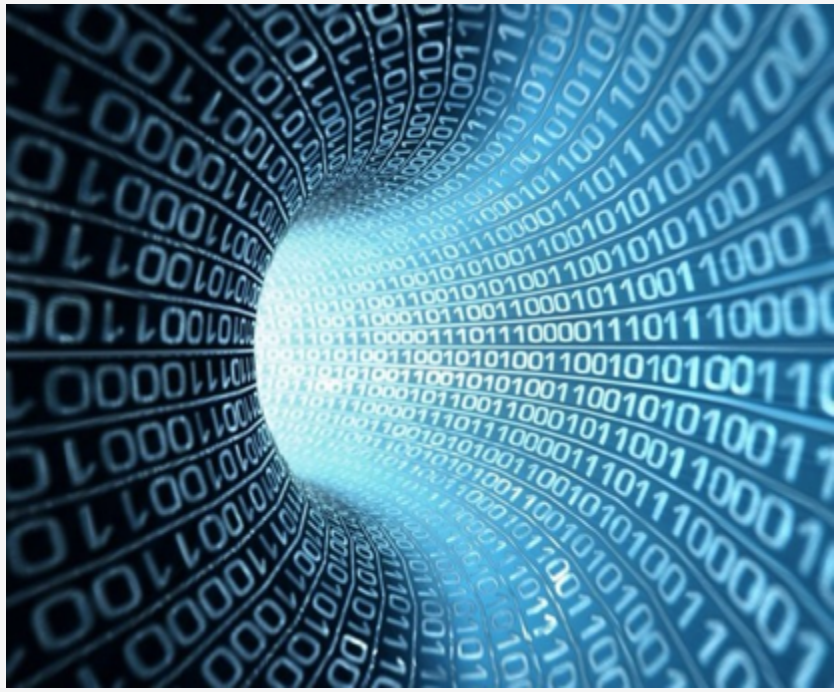
Open Source Solutions



Pig / Hive:

- Cloudera VM
- Cloudera Docker

Open Source Solutions



Cassandra Cluster:

- Datastax distribution
  - free and easy to install
- Datastax community edition

Conclusion

- Big Data courses are very popular.
- Teaching still evolving.
- Infrastructure critical, but expensive and beyond the reach of many schools.
- Open source infrastructure can help get you easily started.
- Students get more experience with installation and administration.