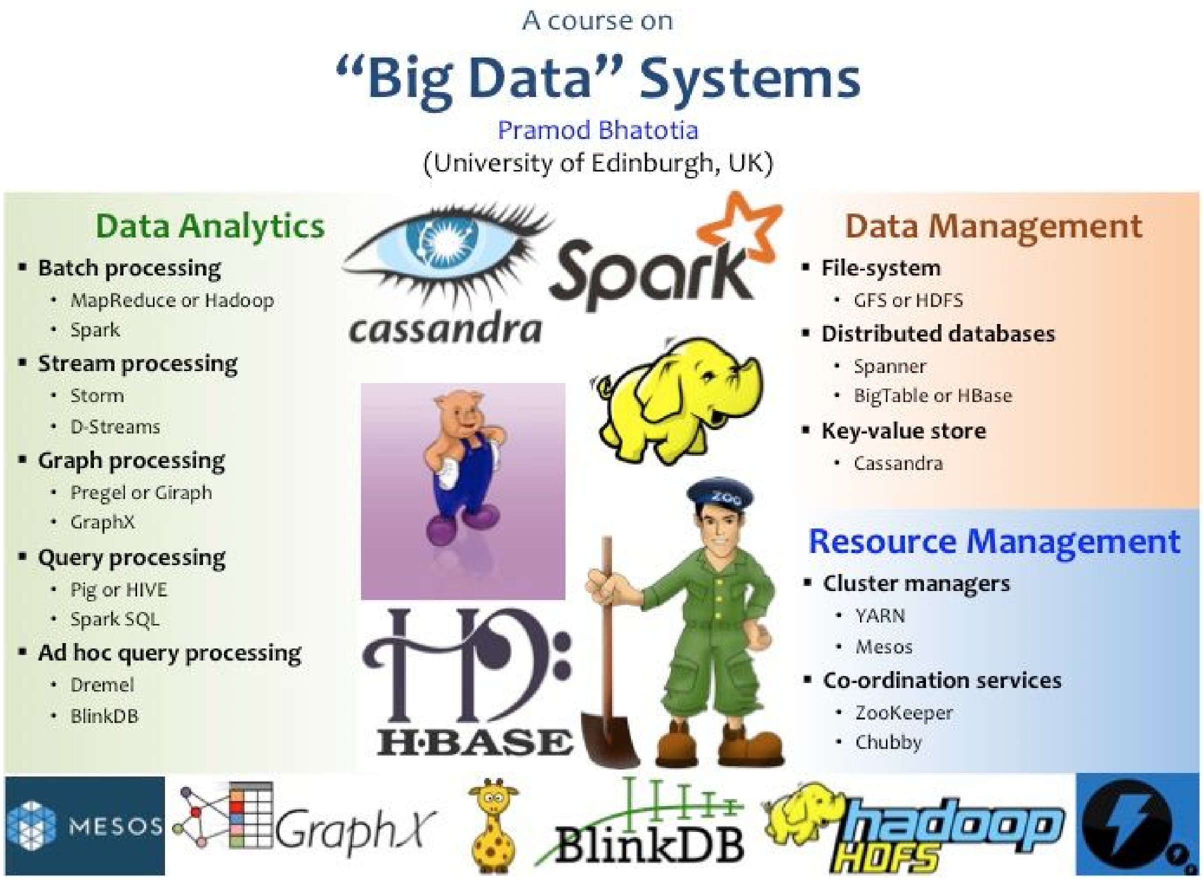

A course on

"Big Data" Systems

Pramod Bhatotia

(University of Edinburgh, UK)

Data Analytics

- Batch processing
  - MapReduce or Hadoop
  - Spark
- Stream processing
  - Storm
  - D-Streams
- Graph processing
  - Pregel or Giraph
  - GraphX
- Query processing
  - Pig or HIVE
  - Spark SQL
- Ad hoc query processing
  - Dremel
  - BlinkDB

Data Management

- File-system
  - GFS or HDFS
- Distributed databases
  - Spanner
  - BigTable or HBase
- Key-value store
  - Cassandra

Resource Management

- Cluster managers
  - YARN
  - Mesos
- Co-ordination services
  - ZooKeeper
  - Chubby

"Big Data" Systems

Course description: The week long course on "Big Data" Systems will mainly cover important & active projects in the broad area of large-scale data management. More specifically, the course will focus on a wide range of open source projects in the area of Big Data. I'll be covering a subset of them, at most five, prioritized based on the interest of students. For each selected project, I plan to cover the following three main points:

1. Why is it important?
2. How to use it?
3. How is it built?

For application programmers, points (1) and (2) are more important; and for systems-sy audience (1) and (3) would be of more interest. I hope that by the end of the course, the students will be in a better position to use these systems and understand what goes behind the scenes.

Format: I'm planning to offer five lectures, each one for roughly 2 hrs. The lecture will be followed by short tutorial exercises for students (1 hr).

Prerequisites: A course on distributed systems (also operating systems, if possible) is required to better understand the concepts, and systems internals. However, in case, the students are mainly targeting to use these systems as an application programmer and don't have the required background then basic programming skills are sufficient.

Since, I don't plan to offer programming tutorials, expert coding skills, *as such*, are not required to participate in the course. But I do expect that students have a fair experience in object oriented programming.

Contact:

My email address: pramod.bhatotia@ed.ac.uk

Tentative course schedule:

Day	**Topic**
Monday	Course overview
	MapReduce/Apache Hadoop and Apache Pig
Tuesday	GFS/HDFS (Google file system)
	BigTable/HBase (Google's Big Table)
Wednesday	Pregel/ Apache Giraph
	(Graph processing)
Thursday	Apache Zookeeper and Chubby
	(Distributed synchronization service)
Friday	Apache Spark and D-Streams (Spark streaming)